STEIN'S LEMMA AND SUBSAMPLING IN LARGE-SCALE OPTIMIZATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Murat A. Erdogdu
October 2019

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Andrea Montanari) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Mohsen Bayati)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Emmanuel Candes)

Approved for the Stanford University Committee on Graduate Studies

Abstract

Statistics and optimization have been closely linked since the very outset. This connection has become more essential lately, mainly because of the recent advances in computational resources, the availability of large amount of data, and the consequent growing interest in statistical and machine learning algorithms. In this dissertation, we will discuss how one can use tools from statistics such as Stein's lemma, subsampling, and shrinkage to design scalable, and efficient optimization algorithms. The focus will be on the large-scale problems where iterative minimization of the empirical risk –or maximization of the log-likelihood– is computationally intractable, i.e., the number of observations n is much larger than the dimension of the parameter p.

In each chapter, we will discuss an efficient estimator or optimization algorithm designed for training a statistical model when the dataset is large, i.e. in the regime $n \gg p \gg 1$. The proposed algorithms have wide applicability to many supervised learning problems such as binary classification with smooth surrogate losses, generalized linear problems in their canonical representation, and M-estimators. The algorithms rely on iterations that are constructed through Stein's lemma, subsampling, and/or shrinkage techniques that achieve quadratic convergence rate, and that are cheaper than any batch optimization method by at least a factor of $\mathcal{O}(p)$. We will discuss theoretical guarantees of the proposed algorithms, along with their convergence behavior in terms of data dimensions. Finally, we will demonstrate their performance on well-known classification and regression problems, through extensive numerical studies on large-scale real datasets, and show that they achieve the highest performance compared to other widely used and specialized algorithms.

To my parents Sultan and Hasan,
and my sister Beril

Acknowledgments

I would like to thank my advisors Mohsen Bayati and Andrea Montanari for their constant support and encouragement throughout my doctoral studies. They have always had their doors open to me when I needed their guidance, despite their busy schedule. There is not a single time that I left their offices without being impressed by their extraordinary scope of knowledge. Their enthusiasm for research has been the main inspiration for me to pursue an academic career.

I would like to thank Emmanuel Candes, Susan Holmes, Lester Mackey, and Ayfer Ozgur for serving in my candidacy, thesis reading, and oral examination committees. I am also grateful to Persi Diaconis for all his help, support and encouragement.

I was fortunate to have wonderful collaborators. I thank Lee H. Dicker, Hera He, Hakan Inan, Ritesh Kolte, Jure Leskovec, Ayfer Ozgur, Anand Rajaraman, Qingyuan Zhao, and Nadia Fawaz for continued discussions, encouragements, and collaborations over the past few years. I have learned too much from them, and I am grateful to their kindness, patience, and support.

I am grateful to all faculty and staff at Department of Statistics at Stanford for creating a fantastic academic environment, and a second home. I thank all my classmates, Bhaswar Bhattacharya, Yunjin Choi, Hera He, Joshua Loftus, Stephen Reid, Milan Shen, Weijie Su, Stefan Wager, Jingshu Wang, Chaojun Wang, and Qingyuan Zhao for many collaborative hours spent on numerous assignments. My office mates Bhaswar, Joshua, and Qingyuan deserve a special mention. I was extremely fortunate to share the office 206 with them.

I am fortunate to have had wonderful friends during my time at Stanford. I am thankful to Cagan Alkan, Kinjal Basu, Selen Bozkurt, Burak Cetin, Berkin Dortdivanlioglu, Burc Eryilmaz, Asli Gundogar, Nilay Gungor, Hakan Inan, Huseyin Inan, Pooja Loftus, Rajarshi Mukherjee, Andres and Muge Notzli, Subhabrata Sen, and Tugce Tasci who made this period of my life very enjoyable.

I am grateful to my wonderful friends with whom I have had countless unforgettable moments. I thank Deniz Badur, Mahmut Ersu, Bulut Esmer, Ozan Gunduz, Selim Kaya, Sahir Nalcaci, Oktay Ozgun, and Yigit Ozpak for the invaluable time we have spent together.

Finally and most importantly, I am grateful to my family, my parents Sultan and Hasan, and my sister Beril. There are no words to describe their support. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Classical regression analysis can be traced back to early 19th century where Carl Friedrich Gauss and Adrien-Marie Legendre invented the method of least squares around the same time [Sti81]. Their objective was to understand the properties of celestial bodies, and they applied least squares to a dataset that is composed of measurements from astronomical objects. Gauss was the first to model the measurement errors using a normal density. He computed the maximum likelihood estimator under the linear model, which led to the celebrated linear regression. Ever since Gauss's pioneering work, traditional statistical methodology follows the same steps. First, a model is proposed by the statistician to explain certain properties of the dataset. Second, the proposed model is trained by carefully formulating it as an optimization problem – for example maximum likelihood estimation by minimizing the negative log-likelihood via Newton-Raphson method. The relationship between statistical methodology and optimization is obviously essential, however – in the broadest sense – statistical estimation techniques have been the primary beneficiary. In this dissertation, we will reverse this arrangement. We will show that optimization algorithms can also immensely benefit from the classical results from statistical estimation theory.

Statistics and optimization have been closely related since the very outset because of the nature of statistical methodology, and the order in which it is carried. This order has not changed much since Gauss, even though there have been tremendous advances in computational resources and the availability of large amount of data which in turn required statisticians to care more about the computational aspects of their models. Datasets with millions of samples and features require careful handling, since off-the-shelf optimization tools are not adequate in general – even simple statistical models may take days to train

using traditional algorithms. Therefore, it is essential for statisticians to devise efficient algorithms to practice their methodology.

Modern data science is facing lots of computational challenges due to availability of tremendous amounts of data, but there are very strong tools in statistics and probability that can turn this surplus to advantage. Having huge amounts of samples allows statistician to reduce an optimization problem that is defined on empirical risk to its population counterpart which is due to the phenomenon called concentration of measure. Then, tools from statistical estimation theory can be used to design fast, efficient, and reliable estimators as well as optimization algorithms. This methodology will be our main tool in this dissertation. First, we will look at the general structure of a statistical learning problem, reduce it to its population version using concentration of measure due to large-scale regime assumption, and then we will design an efficient algorithm on the population version of the problem using techniques from statistical estimation.

1.1 Loss Minimization: Regression and Classification

Many problems in statistics, and machine learning can be formulated as a minimization of the following form

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ f(\beta) \coloneqq \widehat{R}(\beta) + r_\lambda(\beta)\,, \tag{1.1}$$

where the objective function $f : \mathbb{R}^p \to \mathbb{R}$ is the sum of an empirical risk function $\widehat{R} : \mathbb{R}^p \to \mathbb{R}$, and a regularizer $r_\lambda : \mathbb{R}^p \to \mathbb{R}$ at a given penalty level $\lambda \in \mathbb{R}$. The empirical risk function can be written as an average of n functions $f_i : \mathbb{R}^p \to \mathbb{R}$,

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\beta)\,, \tag{1.2}$$

where n denotes the number of samples in the dataset, and p denotes the dimension of the parameter. Functions f_i typically quantify the loss incurred by the sample i. Throughout, we assume that the dataset is very large, i.e., n and p are both large, but n is much larger than p. More specifically, we focus on the regime $n \gg p \gg 1$.

The minimization problem introduced in Equation (1.1) is generally the final step in a statistical estimation method. In statistical learning, each function f_i corresponds to a

measure of misfit or cost of misclassification associated to sample i in the dataset. In order to make this connection more explicit, assume that the dataset consists of n pairs (y_i, x_i) satisfying the classical linear model

$$y_i = \langle x_i, \beta \rangle + \epsilon_i, \quad \text{for } i = 1, 2, ..., n, \tag{1.3}$$

where $y_i \in \mathbb{R}$ represent the response variable which can be continious or binary, $x_i \in \mathbb{R}^p$ represent the set of centered features ($\Sigma_j x_{ij} = 0$), $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$ denotes the Gaussian noise, and $\langle a, b \rangle$ denotes the standard inner product between vectors a and b. Given the above linear model, regularized maximum likelihood estimation can be formulated as in Equation (1.1) where each function corresponds to the negative of the log-likelihood of that particular observation. After getting rid of the redundant terms, maximum likelihood estimator in the above linear model is computed by minimizing the empirical risk in Equation (1.2) where each term in the summation is given as

$$f_i(\beta) = (y_i - \langle x_i, \beta \rangle)^2. \tag{1.4}$$

This is generally referred to as the ordinary least squares or linear regression. More generally, one can use exponential families to model the response variables, which leads to a more flexible setup called generalized linear models.

1.1.1 Generalized Linear Models

Generalized Linear Models (GLMs) play a crucial role in numerous statistical and machine learning problems. GLMs formulate the natural parameter in exponential families as a linear model and provide a miscellaneous framework for statistical methodology and supervised learning tasks. Celebrated examples include linear, logistic, multinomial regressions and applications to graphical models [NB72, MN89, KF09].

We say that the distribution of a random variable $y \in \mathbb{R}$ belongs to an exponential family with natural parameter $\eta \in \mathbb{R}$ if its density can be written as

$$g(y|\eta) = e^{\eta y - \Psi(\eta)} h(y), \tag{1.5}$$

where Ψ is generally referred to as the cumulant generating function, and h is called the carrier density. Let $y_1, y_2, ..., y_n$ be independent observations such that for $i = 1, 2, ..., n$,

the distribution of y_i belongs to an exponential family with natural parameter η_i, i.e., $y_i \sim g(y_i|\eta_i)$. Denoting the vector of natural parameters by $\eta = (\eta_1, ..., \eta_n)^T$, the joint likelihood can be written as

$$g(y_1, y_2, ..., y_n|\eta) = \exp\left\{\sum_{i=1}^{n} [y_i\eta_i - \Psi(\eta_i)]\right\} \prod_{i=1}^{n} h(y_i). \tag{1.6}$$

We will consider the problem of learning the maximum likelihood estimator in the above exponential family framework, where the vector $\eta \in \mathbb{R}^n$ is modeled through the linear relation,

$$\eta = \mathbf{X}\beta, \tag{1.7}$$

for some design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rows $x_i \in \mathbb{R}^p$, and a coefficient vector $\beta \in \mathbb{R}^p$. This formulation is known as Generalized Linear Models (GLMs) under their canonical representation.

The above representation provides a flexible framework for statistical methodology. The cumulant generating function Ψ determines the class of GLMs, i.e., for ordinary least squares (OLS) we use $\Psi(z) = z^2/2$, for logistic regression (LR) we use $\Psi(z) = \log(1 + e^z)$, and for Poisson regression (PR) we use $\Psi(z) = e^z$.

Finding the maximum likelihood estimator in the above formulation is equivalent to minimizing the negative log-likelihood function,

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \Psi(\langle x_i, \beta\rangle) - y_i\langle x_i, \beta\rangle. \tag{1.8}$$

The relation to OLS and LR can be seen much easier by plugging in the corresponding $\Psi(z)$ in Equation (1.8); for example in the case of OLS, compare Equations (1.4) and each summand in Equation (1.8) for $\Psi(z) = z^2/2$. In a classification problem where the dataset has binary response variables $y_i \in \{0, 1\}$, modeling the response with the binomial distribution is equivalent to choosing each function in the empirical risk as

$$f_i(\beta) = \log(1 + \exp(\langle x_i, \beta\rangle)) - y_i\langle x_i, \beta\rangle. \tag{1.9}$$

This leads to the classical logistic regression.

1.1.2 Binary Classification with Smooth Surrogate Loss

The above GLM formulation can be extended to more general settings. For example, let us assume that for $i = 1, 2, ..., n$, the response is binary, $y_i \in \{0, 1\}$. Binary classification with smooth surrogate loss can be described by the following minimization of an empirical risk,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i; q(\langle x_i, \beta \rangle)), \tag{1.10}$$

where ℓ and q are referred to as the loss and the link functions, respectively. Note that we set $f_i(\cdot) = \ell(y_i; q(\langle x_i, \cdot \rangle)$ in Equation (1.2). The loss function ℓ quantifies the cost of misclassification, and there are various loss functions that are used in practice. Examples include log-loss, boosting loss, square loss etc; we will discuss some of them in Chapter 2, i.e, see Table 2.1. As in the previous section, we constrain our analysis to the canonical links. The concept of canonical links for binary classification is introduced by [BSS05], and it is quite similar to the generalized linear problems.

For any given loss function, we define the partial losses $\ell_k(\cdot) = \ell(y = k; \cdot)$ for $k \in \{0, 1\}$. Since we have a binary response variable, we can write any loss in the following format

$$\begin{aligned} \ell(y; q) =& y\ell_1(q) + (1 - y)\ell_0(q), \\ =& y\left(\ell_1(q) - \ell_0(q)\right) + \ell_0(q). \end{aligned}$$

The above formulation is clearly of the form of a generalized linear problem.

1.1.3 Regularization and Constrained Optimization

It is often the case in statistical methodology that a regularization term $r_\lambda(\beta)$ is included to the minimization problem for sparsity and/or shrinkage purposes [FHT01]. This term typically helps to prevent issues such as overfitting, colinearity, by penalizing the complexity of the optimal solution β; examples include restrictions for smoothness, bounds on the vector space norm or constraining the feasible set. There are numerous options for the regularization function r_λ. Below, we only review the most popular ones.

Ridge penalty. The regularization is obtained by penalizing the ℓ_2 norm of the coefficients [HK70, FHT01]. When there are many correlated variables in a linear model, the coefficients may have high variance [FHT01]. As a result, unusually large positive

coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated counterpart. ℓ_2 regularization can prevent this issue.

The minimization problem can be written as an ℓ_2 penalized empirical risk

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; f(\beta) \coloneqq \widehat{R}(\beta) + \frac{\lambda}{2} \|\beta\|_2^2 . \tag{1.11}$$

The penalty level λ is a complexity parameter that controls the amount of shrinkage. The optimal coefficients are shrunk further towards zero for larger values of λ. It should be noted that even though the ridge penalty induces shrinkage towards zero, it does not enforce sparsity on the coefficients.

An equivalent way to introduce ridge penalty is through the following constrained optimization problem

$$\underset{\beta \in \mathcal{C}}{\text{minimize}} \; f(\beta) \coloneqq \widehat{R}(\beta) \quad \text{where } \mathcal{C} = \left\{ \beta \in \mathbb{R}^p \,:\, \|\beta\|_2 \leq \lambda' \right\} . \tag{1.12}$$

There is a one-to-one correspondence between the parameters λ in Equation (1.11) and λ' in Equation (1.12). Hence, one solves either the unconstrained optimization problem in Equation (1.11) or the constrained problem in Equation (1.12). We will discuss these methods in the next section.

Lasso penalty. The lasso penalty is very similar to ridge penalty. In this case, instead of the ℓ_2 norm, one constrains the ℓ_1 norm of the coefficients [Tib96]. In the compressed sensing literature, this type of regularization is commonly referred to as Basis Pursuit De-noising [CDS01]. Lasso regularization dominated the recent statistics literature due its success in variable selection.

The lasso problem can be written as an unconstrained minimization,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; f(\beta) \coloneqq \widehat{R}(\beta) + \lambda \|\beta\|_1 . \tag{1.13}$$

Similar to the ridge penalty, an equivalent unconstrained formulation is

$$\underset{\beta \in \mathcal{C}}{\text{minimize}} \; f(\beta) \coloneqq \widehat{R}(\beta) \quad \text{where } \mathcal{C} = \left\{ \beta \in \mathbb{R}^p \,:\, \|\beta\|_1 \leq \lambda' \right\} . \tag{1.14}$$

Again, there is a one-to-one correspondence between the parameters λ in Equation (1.13) and λ' in Equation (1.14) [FHT01].

Elastic net and ℓ_q penalties. There are many different regularization choices available. For example, a compromise between ridge and lasso is called elastic net. The elastic net penalty can be written as

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ f(\beta) \coloneqq \widehat{R}(\beta) + \alpha\lambda \left\|\beta\right\|_2^2 + (1-\alpha)\lambda \left\|\beta\right\|_1 \quad \text{for} \quad \alpha \in [0,1]. \tag{1.15}$$

The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.

Similar to ℓ_1 and ℓ_2 penalties, one can also use ℓ_q norm for regularization. However, elastic net has significant computational advantages over the ℓ_q penalties [ZH05, FHT01].

Above, we have seen many statistical methods written as an optimization problem. In the next section, we review some popular methods to solve these problems.

1.2 Optimization Methods Used in Statistics

There are numerous optimization algorithms for solving the aforementioned minimization problems. Since the gradient of the objective function is generally non-linear, the optimization method needs be iterative [Bis95, BV04, Nes13]. In the unconstrained case, a standard update rule is given as

$$\beta^{t+1} = \beta^t - \gamma_t \mathbf{Q}^t \nabla_\beta f(\beta^t), \tag{1.16}$$

where γ_t is the step size, and $\mathbf{Q}^t$ is a suitable scaling matrix that provides curvature information. Throughout the introduction, we will focus on the unconstrained case for simplicity. The above iteration – or its projected or proximal version – is also our main focus in Chapters 3 and 4, but with a new approach on how to compute the sequence of scaling matrices $\{\mathbf{Q}^t\}_{t>0}$. We will formulate the problem of finding a scalable $\mathbf{Q}^t$ as an estimation problem and apply a Stein-type lemma, subsampling, and/or shrinkage techniques that provides us with a computationally efficient update rule. It is worth noting that in Chapter 2, we will consider an entirely different technique which will be discussed later.

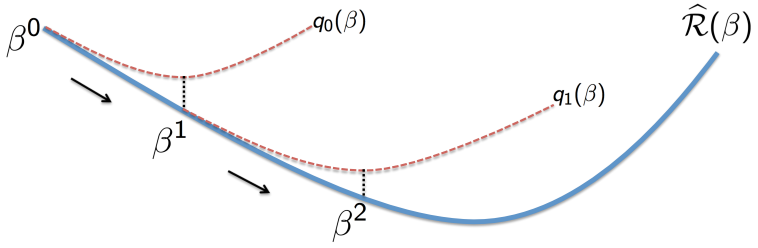

Figure 1.1: Plots show the quadratic approximations performed by an optimization algorithm at each iteration. The quality of the local approximations depend on the quadratic functions $\{q_t\}_{t\geq 0}$.

The standard approach in non-linear continuous optimization is to approximate the objective function $\widehat{R}$ locally with a function $q_t : \mathbb{R}^p \to \mathbb{R}$ at each iteration t, and minimize this local approximation to find the next iteration point. The function sequence $\{q_t\}_{t\geq 0}$ determines the type of the optimization algorithm. Most literature focuses on quadratic approximations of the following form

$$q_t(\beta) = \widehat{R}(\beta^t) + \langle \nabla \widehat{R}(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\gamma_t} \langle \beta - \beta^t, [\mathbf{Q}^t]^{-1}(\beta - \beta^t) \rangle, \tag{1.17}$$

around the current iteration point β^t. This approximation yields the standard update rule as given in Equation (1.16) when minimized over β, i.e.,

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\ q_t(\beta) = \beta^t - \gamma_t \mathbf{Q}^t \nabla \widehat{R}(\beta^t). \tag{1.18}$$

Obviously, the functions q_t that are of the form in Equation (1.17) are quadratic, and they look like second order Taylor series approximations. The quality of the local quadratic approximation is determined by the curvature matrix $\mathbf{Q}^t$. By choosing a suitable curvature (scaling) matrix, one can devise various algorithms.

The iterations of the form Equation (1.16) have been extensively studied in the optimization literature. The case where $\mathbf{Q}^t$ is equal to the identity matrix corresponds to gradient descent (GD) which, under smoothness assumptions, achieves linear convergence rate with $\mathcal{O}(np)$ per-iteration cost. More precisely, gradient descent with ideal step size yields

$$\left\|\beta^{t+1} - \beta_*\right\|_2 \leq \xi_{1,\mathrm{GD}}^t \left\|\beta^t - \beta_*\right\|_2,$$

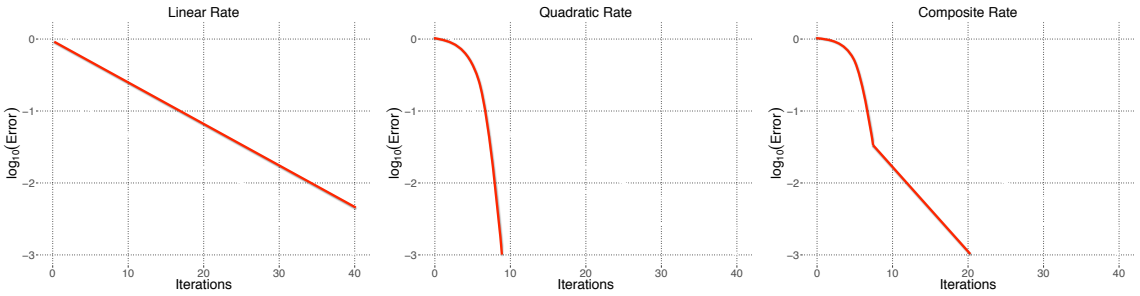

Figure 1.2: Plots show various convergence types. From left to right: Linear convergence obtained by first order methods, quadratic convergence obtained by second order methods, and composite convergence obtained by Newton-Stein method. Error at iteration t is quantified by $\|\hat{\beta}^t - \hat{\beta}_*\|_2$.

where, as $\lim_{t\to\infty} \xi_{1,\mathrm{GD}}^t = 1 - (\lambda_p^*/\lambda_1^*)$, and λ_i^* is the i-th largest eigenvalue of the Hessian of $f(\beta)$ at the true minimizer β_* of $\widehat{R}$.

Second order methods such as Newton method (NM) or Newton-Raphson and natural gradient descent (NGD) [Ama98] can be recovered by taking $\mathbf{Q}^t$ to be the inverse Hessian and the Fisher information evaluated at the current iterate, respectively. Such methods may achieve quadratic convergence rates with $\mathcal{O}(np^2 + p^3)$ per-iteration cost [Bis95, Nes13]. In particular, for t large enough, Newton method yields

$$\left\|\beta^{t+1} - \beta_*\right\|_2 \leq \xi_{2,\mathrm{NM}}^t \left\|\beta^t - \beta_*\right\|_2^2,$$

and it is insensitive to the condition number of the Hessian. However, when the number of samples grows large, computation of $\mathbf{Q}^t$ becomes extremely expensive. First two plots in Figure 1.2 depicts linear, quadratic convergence rates, respectively.

It is quite common in statistical learning that each step of Newton-Raphson method is formulated as a weighted least squares problem [FHT10]. For example, in order to compute the maximum likelihood estimator in generalized linear problems, one can solve a weighted least squares problem at each iteration. This formulation is commonly referred to as iteratively re-weighted least squares (IRLS). We emphasize that IRLS and standard Newton-Raphson methods are equivalent for GLMs under their canonical representation [MN89, FHT10], and we will use the latter to refer to this algorithm throughout this dissertation.

A popular line of research tries to construct the matrix $\mathbf{Q}^t$ in a way that the update

is computationally feasible, yet still provides sufficient second order information. This can be accomplished by estimating the Hessian with a computationally efficient matrix. Such attempts resulted in Quasi-Newton methods, in which only gradients and iterates are used in the construction of matrix $\mathbf{Q}^t$, resulting in an efficient update at each step t [Nes13, Bis95]. These algorithms are based on the Quasi-Newton relation which is given as,

$$\mathbf{Q}^{t+1}(\nabla f(\beta^{t+1}) - \nabla f(\beta^t)) = \beta^{t+1} - \beta^t. \tag{1.19}$$

Denoting by $\Delta\mathbf{Q}^t = \mathbf{Q}^{t+1} - \mathbf{Q}^t$, $g^t = \nabla f(\beta^{t+1}) - \nabla f(\beta^t)$, and $d^t = \beta^{t+1} - \beta^t$, several popular Quasi-Newton updates can be written as below [Nes13, Bis95].

- Rank-one correction scheme:

$$\Delta\mathbf{Q}^t = \frac{(d^t - \mathbf{Q}^t g^t)(d^t - \mathbf{Q}^t g^t)^T}{\langle d^t - \mathbf{Q}^t g^t, g^t\rangle}. \tag{1.20}$$

- Davidon-Fletcher-Powell scheme:

$$\Delta\mathbf{Q}^t = \frac{d^t[d^t]^T}{\langle d^t, g^t\rangle} - \frac{\mathbf{Q}^t g^t[g^t]^T\mathbf{Q}^t}{\langle \mathbf{Q}^t g^t, g^t\rangle}. \tag{1.21}$$

- Broyden-Fletcher-Goldfarb-Shanno (BFGS) scheme:

$$\Delta\mathbf{Q}^t = \frac{\mathbf{Q}^t g^t[d^t]^T + d^t[g^t]^T\mathbf{Q}^t}{\langle \mathbf{Q}^t g^t, g^t\rangle} - \alpha^t\frac{\mathbf{Q}^t g^t[g^t]^T\mathbf{Q}^t}{\langle \mathbf{Q}^t g^t, g^t\rangle}, \tag{1.22}$$

  where $\alpha_t = 1 + \langle g^t, d^t\rangle/\langle \mathbf{Q}^t g^t, g^t\rangle$.

From the computational point of view, BFGS is considered as the most stable scheme [Bro70, Fle70, Gol70, Sha70, Nes13, Bis95]. For completeness, we note that it requires $\mathcal{O}(np + p^2)$ per-iteration cost [Bis95, Nes13].

Finally, another popular approach in large-scale optimization is to use subsampling techniques where scaling matrix $\mathbf{Q}^t$ is constructed through randomly selected set of data points [Mar10, BCNN11, EM15]. At iteration t, the scaling matrix $\mathbf{Q}^t$ is computed using a subset of the samples which reduces the cost of computing the Hessian substantially in the regime we consider. That is, the cost of computing the full Hessian is generally $\mathcal{O}\left(np^2\right)$ whereas using a subsample $S \subset [n]$, this can be reduced to $\mathcal{O}\left(|S|p^2\right)$. Subsampling has been widely used in both first and second order methods, but was not as well studied in the

second order case for which the scaling matrix is approximated. In particular, theoretical guarantees were still missing. We will discuss this in detail in Chapter 4, and see that they achieve composite convergence rate which can be seen from the right plot in Figure 1.2. We will discuss this type of convergence in detail in Chapters 3 and 4.

1.3 Computational Challenges in Large-Scale Optimization Problems

The main challenge in problems of the form Equation (1.1) is to balance the trade-off between the per-iteration cost of an optimization algorithm and its convergence rate. First order methods enjoy cheaper per-iteration cost of $\mathcal{O}(np)$ (n is the sample size, p is the dimension), but they achieve locally linear convergence rate which is considered as slow in batch optimization. On the other hand, second order methods enjoy quadratic convergence rate, yet their per-iteration cost is $\mathcal{O}\left(np^2\right)$. In the large-scale regime, where $n \gg p \gg 1$, per-iteration cost of $\mathcal{O}\left(np^2\right)$ is not affordable. Below, we describe where these main computational issues stem from.

Cost of computing Hessian. Given a problem of the form in Equation (1.1), standard Newton update requires the computation of the Hessian which is of the form

$$\boldsymbol{\nabla}^2 f(\beta) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\nabla}^2 f_i(\beta). \tag{1.23}$$

Notice that the Hessian is an average of n matrices, and each one belongs to $\mathbb{R}^{p\times p}$. Unless there is a special structure in the problem, one needs to compute the matrix $\boldsymbol{\nabla}^2 f$ in order to use Newton method, and enjoy quadratic convergence. Since we average over n matrices where each has p^2 entries, the cost of computing the Hessian becomes $\mathcal{O}\left(np^2\right)$. In the large-scale problems, namely $n \gg p \gg 1$, this is the main bottleneck for the second order optimization methods.

Cost of computing the inverse of the Hessian. For large-scale problems, when the dimension of the feature vectors p is large, computing the inverse of a $p \times p$ matrix may be extensive. Standard inversion algorithms can perform this operation with $\mathcal{O}\left(p^3\right)$ computation (or a little better than p^3). In this regime, classical approach is

to avoid inversion by solving the following linear system of equations

$$-\boldsymbol{\nabla}^2 f(\beta)d = \nabla f(\beta), \tag{1.24}$$

and taking a step towards the direction of d. The above system may be solved exactly by various factorization methods or approximately by using iterative methods [HS52, PS75]. Iterative methods are proven to be effective in the large-scale regime, but these methods typically specialize on certain types of equations. For the problems we consider, we will assume that the Hessian is positive semidefinite, hence the Cholesky factorization and conjugate gradient methods are applicable. It is important to highlight that these methods work well when p is too large, where matrix inversion is not affordable. In terms of approximations, conjugate gradient methods work very well, but early stopping may result in imprecise directions, and late stopping may cause serious numerical issues due to the iterations in Lanczos process [HS52, PS75]. For moderate values of p, exact inversion of the Hessian is affordable.

Fluctuations due to subsampling. Subsampled Newton methods have been proven to be very useful in practice [Mar10, BCNN11, EM15]. A key challenge is that the subsampled Hessian is close to the actual Hessian along the directions corresponding to large eigenvalues (large curvature directions in $f(\beta)$), but is a poor approximation in the directions corresponding to small eigenvalues (flatter directions in $f(\beta)$). This yields poor estimation of Hessian, and as a result, the resulting subsampled Newton algorithm may have undesirable convergence, or even fail to converge. The same issue is commonly encountered in statistical estimation theory as well. Especially in covariance estimation, it is well-known that the performance of various estimators can be significantly improved by simple procedures such as shrinkage and/or thresholding [CCS10, DGJ13, GD14]. To this extent, we will use a specialized low-rank approximation as the important second order information is generally contained in the largest few eigenvalues/vectors of the Hessian.

The computational challenges introduce various limitations in the practical settings [EM15, EF15, EFM15]. Using certain tools from statistics and probability theory such as Stein's lemma, zero-biased transformations, subsampling and shrinkage techniques we will remedy some of these issues.

1.4 Summary of Contributions

We propose efficient estimators and/or optimization algorithms for large-scale problems by assuming random design, and applying Stein's lemma, subsampling, and/or shrinkage techniques. Each chapter will introduce and discuss a different optimization approach, and study its theoretical properties.

1.4.1 Scalable First Order Stein Approximations for GLMs

In Chapter 2, we take an unconventional approach for minimizing Equation (2.1), based on an identity that is well-known in some areas of statistics, but appears to have received relatively little attention for its computational implications in large-scale problems. Let β^{pop} denote the true minimizer of the population version of the risk function given in Equation (1.2), and let β^{ols} denote the corresponding ordinary least squares (OLS) coefficients defined as $\beta^{\mathsf{ols}} = \mathbb{E}\left[xx^T\right]^{-1}\mathbb{E}\left[xy\right]$. Then, under certain random predictor (design) models,

$$\beta^{\mathsf{pop}} \propto \beta^{\mathsf{ols}}. \tag{1.25}$$

For logistic regression with Gaussian design (which is equivalent to Fisher's discriminant analysis), Equation (1.25) was noted by Fisher in the 1930s [Fis36]; a more general formulation for models with Gaussian design is given in [Bri82]. The relationship Equation (1.25) suggests that if the constant of proportionality is known, then β^{pop} can be estimated by computing the OLS estimator, which may be substantially simpler than minimizing the empirical risk. In fact, in some applications like binary classification, it may not be necessary to find the constant of proportionality in Equation (1.25). Our work in this chapter builds on this idea.

Our contributions can be summarized as follows.

- We show that β^{pop} is approximately proportional to β^{ols} in the random design setting, regardless of the covariate (predictor) distribution. That is, we prove

$$\left\|\beta^{\mathsf{pop}} - c_\Psi \times \beta^{\mathsf{ols}}\right\|_\infty \lesssim \frac{1}{p},$$

  for some $c_\Psi \in \mathbb{R}$ which depends on the non-linearity Ψ. Our generalization uses zero-bias transformations [GR97]. We also show that the above relation still holds under certain types of regularization.

- We design a computationally efficient estimator for β^{pop} by first estimating the OLS coefficients, and then estimating the proportionality constant c_Ψ via line search. We refer to the resulting estimator as the Scaled Least Squares (SLS) estimator and denote it by $\hat{\beta}^{\mathsf{sls}}$. After estimating the OLS coefficients, the second step of our algorithm involves finding a root of a real valued function; this can be accomplished using iterative methods with up to a quadratic convergence rate and only $\mathcal{O}(n)$ per-iteration cost. This is cheaper than the classical batch methods mentioned above by at least a factor of $\mathcal{O}(p)$.

- For random design with sub-Gaussian predictors, we show that

$$\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_\infty \lesssim \frac{1}{p} + \sqrt{\frac{p}{n/\log(n)}}.$$

  This bound characterizes the performance of the proposed estimator in terms of data dimensions, and justifies the use of the algorithm in the regime $n \gg p \gg 1$.

- We demonstrate how to transform a binary classification problem with smooth surrogate loss into a generalized linear problem, and how our methods can be applied to obtain a computationally efficient optimization scheme. We further discuss the canonicalization of the square loss, which may be of independent interest to non-convex optimization community.

- We propose a scalable algorithm for converting one generalized linear problem to another by exploiting the proportionality relation Equation (1.25). The proposed algorithm requires only $\mathcal{O}(n)$ per each iteration, with no additional cost.

- We study the statistical and computational performance of $\hat{\beta}^{\mathsf{sls}}$, and compare it to that of the empirical risk minimizer (using several well-known implementations), on a variety of large-scale datasets.

1.4.2 Newton-Stein Method: A New Second Order Method

In Chapter 3, we focus on how to solve the maximum likelihood problem efficiently in the GLM setting when the number of observations n is much larger than the dimension of the coefficient vector p, i.e., $n \gg p \gg 1$. The optimization algorithms for solving the GLM problem were discussed in Section 1.1.1, where the objective function was the negative of

the log-likelihood, however only a few of these algorithms can utilize the special structure of GLMs. In this chapter, we propose an algorithm which takes a Newton step that utilizes this special structure by using a Stein-type lemma [Ste81] along with subsampling techniques. It attains fast convergence rates with low per-iteration cost. The proposed algorithm is called Newton-Stein method which we abbreviate as NewSt. Our contributions in this chapter can be summarized as follows:

- We recast the problem of constructing a scaling matrix as an estimation problem and apply a Stein-type lemma along with subsampling techniques to form a computationally feasible $\mathbf{Q}$.
- Newton-Stein method allows further improvements through eigenvalue shrinkage, eigenvalue thresholding, subsampling and various other techniques that are available for covariance estimation.
- Excessive per-iteration cost of $\mathcal{O}(np^2+p^3)$ of Newton method is replaced by $\mathcal{O}(np+p^2)$ per-iteration cost, and a one-time $\mathcal{O}(|S|p^2)$ cost, where $|S|$ is the subsample size.
- Assuming that the rows of the design matrix are i.i.d. and have bounded support (or sub-Gaussian), and denoting the iterates of Newton-Stein method by $\{\hat{\beta}^t\}_t$, we prove a bound of the form

$$\left\|\hat{\beta}^{t+1}-\beta_*\right\|_2 \leq \tau_1\left\|\hat{\beta}^t-\beta_*\right\|_2+\tau_2\left\|\hat{\beta}^t-\beta_*\right\|_2^2, \tag{1.26}$$

  where β_* is the true minimizer and τ_1, τ_2 are the convergence coefficients, and they are deterministic. The above bound implies that the local convergence starts with a quadratic phase and transitions into linear as the iterate gets closer to the true minimizer. We further establish a global convergence result of Newton-Stein method coupled with a line search algorithm.
- We demonstrate the performance of Newton-Stein method on real and synthetic data sets by comparing it to commonly used optimization algorithms.

1.4.3 Convergence Rates of Subsampled Newton Methods

In Chapter 4, we study the local convergence properties of subsampled Newton methods. The algorithms presented in this chapter are more general than the algorithms discussed in

Chapters 2 and 3 in the sense that they are applicable to a broader class of machine learning problems. As previously mentioned in Section 1.2, subsampled Newton methods are commonly used in practice with numerous applications [Mar10, BCNN11, VP12, EM15, Erd15]. If the subsampled Hessian is close to the true Hessian, these methods can approach Newton method in terms of convergence rate, nevertheless, they enjoy much smaller complexity per update. No convergence rate analysis is available for these methods; this analysis is the main contribution of this chapter. Relying on random matrix theory, and empirical risk theory, we derive the local convergence properties of subsampled Newton methods. To the best of our knowledge, the best result in this direction is proven in [BCNN11] that estabilishes asymptotic convergence without quantitative bounds (by the theory from [GNS09]).

- We show that subsampled Newton methods enjoy a composite convergence rate: quadratic at start and linear near the minimizer, as illustrated in the right plot in Figure 1.2. Formally, we prove a bound of the form

$$\|\hat{\beta}^{t+1} - \hat{\beta}_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \hat{\beta}_*\|_2 + \xi_2^t \|\hat{\beta}^t - \hat{\beta}_*\|_2^2, \tag{1.27}$$

  with coefficients that are explicitly given (and are computable from data). Note that this is similar to the bound given for Newton-Stein method, but this time the coefficients are random variables.

- We propose a new subsampled Newton method that relies on subsampling and eigenvalue thresholding. We call our algorithm NewSamp and show that it enjoys composite convergence rate as given in Equation (1.27).

- The asymptiotic behavior of the linear convergence coefficient is $\lim_{t\to\infty} \xi_1^t = 1 - (\lambda_p^*/\lambda_{r+1}^*) + \delta$, for δ small. The condition number $(\lambda_1^*/\lambda_p^*)$ which controls the convergence of gradient decent, has been replaced by the milder $(\lambda_{r+1}^*/\lambda_p^*)$. For datasets with strong spectral features, this can be a large improvement.

- The complexity per iteration of NewSamp is $\mathcal{O}(np + |S|p^2)$ with $|S|$ the sample size. In the large-scale regime where $n \gg p$, if a subsample size of $\mathcal{O}(n/p)$ is used, cost at each iteration becomes $\mathcal{O}(np)$. This is the per-iteration cost of gradient descent.

- Finally, we demonstrate the performance of NewSamp on four datasets, and compare it to the well-known optimization methods.

1.5 Organization, Published Materials, and Acknowledgments

This dissertation is organized as follows. Chapter 2 introduces a new estimator by applying Stein's lemma to the gradient, and using the subsampling technique. Contents of this chapter are based on the papers [EBD16b, EBD16a]. Chapter 3 proposes a second order method by applying Stein's lemma and subsampling techniques to estimate the expectation of the Hessian. Contents of this chapter are based on the papers [Erd15, Erd16]. In Chapter 4, we analyze the convergence rates of subsampled Newton methods, and propose a new second order method with the additional improvements via shrinkage techniques. Contents of this chapter are based on the paper [EM15]. Finally, we conclude with a brief discussion in Chapter 5.

The papers [EBD16b, EBD16a] are joint works with Mohsen Bayati, and Lee H. Dicker. The paper [EM15] is joint work with Andrea Montanari. The papers [Erd15, Erd16] are my own work, though I am grateful to Mohsen Bayati and Andrea Montanari for stimulating conversations on the topics of these works.

Chapter 2

First Order Stein Approximations to Gradient

Contents of this chapter are based on the papers [EBD16b, EBD16a]. In this chapter, we show that under random sub-Gaussian design, the true minimizer of the population risk in a generalized linear problem is approximately proportional to the corresponding ordinary least squares (OLS) estimator. This is obtained by applying a variant of Stein's lemma known as zero-bias transformation to the gradient. Using this relation, we design an algorithm that achieves almost the same accuracy as the empirical risk minimizer through iterations that attain up to a quadratic convergence rate, and that are cheaper than any batch optimization algorithm by at least a factor of $\mathcal{O}(p)$. During this project, Murat A. Erdogdu was partially supported by National Science Foundation grant CMMI:1554140.

2.1 Introduction

Consider the following stochastic optimization problem over the population risk

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ R(\beta) \coloneqq \mathbb{E}\left[\Psi\left(\langle x,\beta\rangle\right) - y\langle x,\beta\rangle\right], \tag{2.1}$$

where $\Psi:\mathbb{R}\to\mathbb{R}$ is a non-linear function, $y\in\mathcal{Y}\subset\mathbb{R}$ denotes the response variable, $x\in\mathcal{X}\subset\mathbb{R}^p$ denotes the predictor (or covariate), and the expectation is over the joint distribution of (y,x). The above minimization is called a generalized linear problem in its canonical representation, and it is commonly encountered in the statistical learning. Celebrated

examples include binary classification with smooth surrogate losses [BSS05, RW10], and generalized linear models (GLMs) such as Poisson regression, logistic regression, ordinary least squares, multinomial regression and many applications involving graphical models [NB72, MN89, WJ08, KF09]. These methods play a crucial role in numerous machine learning and statistics problems, and provide a miscellaneous framework for many regression and classification tasks.

The exact minimization of the stochastic optimization problem given as Equation (2.1), requires the knowledge of the underlying distribution of the variables (y, x). In practice, however, the joint distribution is not available. Therefore, after observing n independent data points (y_i, x_i), the standard approach is to minimize the empirical risk approximation given as

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \widehat{R}(\beta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \Psi\left(\langle x_i, \beta \rangle\right) - y_i \langle x_i, \beta \rangle \,. \tag{2.2}$$

In the case of GLMs, the empirical risk minimization given in Equation (2.2) is called the maximum likelihood estimation, whereas in the case of binary classification, it is generally referred to as surrogate loss minimization. Due to non-linear structure of the optimization task given in Equation (2.2), for both problems, the minimization of the empirical risk requires iterative methods that we have reviewed in Section 1.2. Regardless of the problem formulation, the most commonly used optimization method is the Newton-Raphson method, which may be viewed as a reweighted least squares algorithm [MN89, BSS05]. This method uses a second-order approximation to benefit from the curvature of the log-likelihood and achieves locally quadratic convergence. A drawback of this approach is its excessive per-iteration cost of $\mathcal{O}(np^2)$. On the other hand, first-order approximation yields the gradient descent algorithm, which attains a linear convergence rate with $\mathcal{O}(np)$ per-iteration cost. Although its convergence rate is slow compared to that of the second-order methods, its modest per-iteration cost makes it practical for large-scale problems. In the regime $n \gg p$, another popular optimization technique is the class of Quasi-Newton methods [Bis95, Nes13], which can attain a per-iteration cost of $\mathcal{O}(np)$, and the convergence rate is locally super-linear; a well-known member of this class of methods is the BFGS algorithm [Bro70, Fle70, Gol70, Sha70].

In this chapter, we consider a different approach, based on an identity that is well-known in some areas of statistics, but appears to have received relatively little attention

for its computational implications in large-scale problems. Assume that β^{pop} denotes the true minimizer of the population risk given in Equation (2.1), and let β^{ols} denote the corresponding ordinary least squares (OLS) coefficients defined as $\beta^{\mathsf{ols}} = \mathbb{E}\left[xx^T\right]^{-1}\mathbb{E}\left[xy\right]$. Then, under certain random design models, we have

$$\beta^{\mathsf{pop}} \propto \beta^{\mathsf{ols}}. \tag{2.3}$$

It is well-known that for logistic regression with Gaussian design (which is equivalent to Fisher's discriminant analysis), Equation (2.3) was noted by Fisher in the 1930s [Fis36]; a more general formulation for models with Gaussian design is given in [Bri82]. The relationship Equation (2.3) suggests that if the constant of proportionality is known, then β^{pop} can be estimated by computing the OLS estimator, which may be substantially simpler than minimizing the empirical risk. In fact, in some applications like binary classification, it may not be necessary to find the constant of proportionality in Equation (2.3). Our work in this chapter builds on this idea.

The rest of the chapter is organized as follows: Section 2.1.1 surveys the related work and Section 2.2 introduces the required background and the notation. In Section 2.3, we provide the intuition behind the relationship Equation (2.3), which are based on exact calculations for the Gaussian design setting. In Section 2.4, we propose our algorithm and discuss its computational properties. Theoretical results are given in Section 2.5. In Section 2.6, we propose an algorithm to convert one GLM type to another. We discuss how a binary classification problem can be cast as a generalized linear problem in Section 2.7, and in Section 2.8 we propose a method to canonicalize the square loss. Section 2.9 provides a thorough comparison between the proposed algorithm and other existing methods. Finally, we conclude this chapter with a brief discussion in Section 2.11.

2.1.1 Related work

As mentioned in Section 2.1, the relationship Equation (2.3) is well-known in several forms in statistics. Brillinger [Bri82] derived Equation (2.3) for models with Gaussian predictors using Stein's lemma. Li & Duan [LD89] studied model misspecification problems in statistics and derived Equation (2.3) when the predictor distribution has linear conditional means (this is a slight generalization of Gaussian predictors). The relation Equation (2.3) has led to various techniques for dimension reduction [Li91, LD09], and more recently, it has been

studied by [PV16, TAH15] in the context of compressed sensing. It has been shown that the standard lasso estimator may be very effective when used in models where the relationship between the expected response and the signal is nonlinear, and the predictors (i.e. the design or sensing matrix) are Gaussian. A common theme for all of this previous work is that it focuses solely on settings where Equation (2.3) holds exactly and the predictors are Gaussian (or, in the case of [LD89], very nearly Gaussian). Random design assumptions and Stein's have been a frequent theme in many recent works [BEM13, Erd17, DE16]. Two key novelties of the present chapter are (i) our focus on the computational benefits following from Equation (2.3) for large scale problems with $n \gg p \gg 1$; and (ii) our rigorous finite sample analysis of models with non-Gaussian predictors, where Equation (2.3) is shown to be approximately valid. To the best of our knowledge, the present chapter and its earlier version [EBD16b] are the first to consider the relation Equation (2.3) in the context of optimization.

2.2 Preliminaries and Notation

We assume a random design setting, where the observed data consists of n random iid pairs (y_1, x_1), (y_2, x_2), ..., (y_n, x_n); $y_i \in \mathcal{Y} \subset \mathbb{R}$ is the response variable and $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathcal{X} \subset \mathbb{R}^p$ is the vector of predictors or covariates. We focus on problems where the minimization Equation (2.1) is desirable, but we do not need to assume that (y_i, x_i) are actually drawn from a particular distribution or the corresponding statistical model (i.e. we allow for model misspecification).

$$\beta^{\mathsf{pop}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \mathbb{E}\left[\Psi(\langle x_i, \beta \rangle) - y_i \langle x_i, \beta \rangle\right]. \tag{2.4}$$

While we make no assumptions on Ψ beyond smoothness, note that when the optimization problem is GLM, and Ψ is the cumulant generating function for $y_i \mid x_i$, then the problem reduces to the standard GLM with canonical link and regression parameters β^{pop} [MN89]. Examples of GLMs in this form include logistic regression with $\Psi(w) = \log\{1+e^w\}$, Poisson regression with $\Psi(w) = e^w$, and linear regression (least squares) with $\Psi(w) = w^2/2$.

Our objective is to find a computationally efficient estimator for β^{pop}. The alternative estimator for β^{pop} proposed in this chapter is related to the OLS coefficient vector, which is defined by $\beta^{\mathsf{ols}} := \mathbb{E}[x_i x_i^T]^{-1}\mathbb{E}[x_i y_i]$; the corresponding OLS estimator is $\hat{\beta}^{\mathsf{ols}} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$, where $\mathbf{X} = (x_1, \ldots, x_n)^T$ is the $n \times p$ design matrix and $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$.

Additionally, throughout the text we let $[m]=\{1,2,...,m\}$, for positive integers m, and we denote the size of a set S by $|S|$. The m-th derivative of a function $g:\mathbb{R}\to\mathbb{R}$ is denoted by $g^{(m)}$. For a vector $u\in\mathbb{R}^p$ and a $n\times p$ matrix $\mathbf{U}$, we let $\|u\|_q$ and $\|\mathbf{U}\|_q$ denote the ℓ_q-vector and -operator norms, respectively. If $S\subseteq[n]$, let $\mathbf{U}_S$ denote the $|S|\times p$ matrix obtained from $\mathbf{U}$ by extracting the rows that are indexed by S. For a symmetric matrix $\mathbf{M}\in\mathbb{R}^{p\times p}$, $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the maximum and minimum eigenvalues, respectively, and $\rho_k(\mathbf{M})$ denotes the condition number of $\mathbf{M}$ with respect to k-norm. We denote by N_q the q-variate normal distribution, and all expectations are over all randomness inside the brackets. Finally, we use $a\lesssim b$ and $a\leq\mathcal{O}(b)$ interchangeably, whichever is convenient (where $\mathcal{O}(\cdot)$ refers to the big O notation).

2.3 From OLS to True Minimizer: Gaussian Case

To motivate our methodology, we assume in this section that the covariates are multivariate normal, as in [Bri82]. These distributional assumptions will be relaxed in Section 2.5.

Proposition 2.3.1. *Assume that the covariates are multivariate normal with mean 0 and covariance matrix $\mathbf{\Sigma}$, i.e. $x_i\sim\mathsf{N}_p(0,\mathbf{\Sigma})$. Then β^{pop} can be written as*

$$\beta^{\mathsf{pop}}=c_\Psi\times\beta^{\mathsf{ols}}, \tag{2.5}$$

where $c_\Psi\in\mathbb{R}$ is the fixed point of the mapping

$$z\to\mathbb{E}\left[\Psi^{(2)}(\langle x_i,\beta^{\mathsf{ols}}\rangle z)\right]^{-1}. \tag{2.6}$$

Proof of Proposition 2.3.1. The optimal point in the optimization problem Equation (2.4), has to satisfy the following normal equations,

$$\mathbb{E}\left[yx_i\right]=\mathbb{E}\left[x_i\Psi^{(1)}(\langle x_i,\beta\rangle)\right]. \tag{2.7}$$

Now, denote by $\phi(x\mid\mathbf{\Sigma})$ the multivariate normal density with mean 0 and covariance matrix $\mathbf{\Sigma}$. We recall the well-known property of Gaussian density $\mathrm{d}\phi(x\mid\mathbf{\Sigma})/\mathrm{d}x=-\mathbf{\Sigma}^{-1}x\phi(x\mid\mathbf{\Sigma})$.

Using this and integration by parts on the right hand side of the above equation, we obtain

$$\begin{aligned}\mathbb{E}\left[x_i\Psi^{(1)}(\langle x_i,\beta\rangle)\right] &= \int x\Psi^{(1)}(\langle x,\beta\rangle)\phi(x \mid \mathbf{\Sigma})\ \mathrm{d}x, \\ &= \mathbf{\Sigma}\beta\ \underbrace{\mathbb{E}\big[\Psi^{(2)}(\langle x_i,\beta\rangle)\big]}_{\in\ \mathbb{R}},\end{aligned} \tag{2.8}$$

which is basically the Stein's lemma. Combining this with the normal equations Equation (2.7) and multiplying both side with $\mathbf{\Sigma}^{-1}$, we obtain the desired result. □

Proposition 2.3.1 and its proof provide the main intuition behind our proposed method. Observe that in our derivation, we only worked with the right hand side of the normal equations Equation (2.7) which does not depend on the response variable y_i. Therefore, the equivalence will hold regardless of the joint distribution of (y_i, x_i). This is the main difference from the proof of [Bri82] where y_i is assumed to follow a single index model. In Section 2.5, where we extend the method to non-Gaussian predictors, the identity Equation (2.8) is generalized via the zero-bias transformations [GR97].

2.3.1 Regularization

A version of Proposition 2.3.1 incorporating regularization — an important tool for datasets where p is large relative to n or the predictors are highly collinear — is also possible, as outlined briefly in this section. We focus on ℓ^2-regularization (ridge regression) in this section; some connections with lasso (ℓ^1-regularization) are discussed in Section 2.5 and Corollary 2.5.2.

For $\lambda \geq 0$, define the ℓ_2-regularized empirical risk minimizer,

$$\beta_\lambda^{\mathsf{pop}} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathbb{E}\left[\Psi(\langle x_i,\beta\rangle) - y_i\langle x_i,\beta\rangle\right] + \frac{\lambda}{2}\left\|\beta\right\|_2^2 \tag{2.9}$$

and the corresponding ℓ^2-regularized OLS coefficients $\beta_\lambda^{\mathsf{ols}} = \left(\mathbb{E}\left[x_i x_i^T\right] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}\left[x_i y_i\right]$ (so $\beta^{\mathsf{pop}} = \beta_0^{\mathsf{pop}}$ and $\beta^{\mathsf{ols}} = \beta_0^{\mathsf{ols}}$). The same argument as above implies that

$$\beta_\lambda^{\mathsf{pop}} = c_\Psi \times \beta_\gamma^{\mathsf{ols}}, \quad \text{where} \quad \gamma = \lambda c_\Psi. \tag{2.10}$$

This suggests that the ordinary ridge regression for the linear model can be used to estimate the ℓ^2-regularized empirical risk minimizer $\beta_\lambda^{\mathsf{pop}}$. Further pursuing these ideas for problems

Algorithm 1 SLS: Scaled Least Squares Estimator

Input: Data $(y_i, x_i)_{i=1}^n$

Step 1. Compute the least squares estimator: $\hat{\beta}^{\mathsf{ols}}$ and $\hat{y} = \mathbf{X}\hat{\beta}^{\mathsf{ols}}$.

For a subsampling based OLS estimator, let $S \subset [n]$ be a random subset and take $\hat{\beta}^{\mathsf{ols}} = \frac{|S|}{n}(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}^T y$.

Step 2. Solve the following equation for $c \in \mathbb{R}$: $1 = \frac{c}{n}\sum_{i=1}^n \Psi^{(2)}(c\,\hat{y}_i)$.

Use Newton root-finding method:

Initialize c;

Repeat until convergence:

$$c \leftarrow c - \frac{c\,\frac{1}{n}\sum_{i=1}^n \Psi^{(2)}(c\,\hat{y}_i) - 1}{\frac{1}{n}\sum_{i=1}^n \left\{\Psi^{(2)}(c\,\hat{y}_i) + c\,\hat{y}_i\Psi^{(3)}(c\,\hat{y}_i)\right\}}.$$

Output: $\hat{\beta}^{\mathsf{sls}} = c \times \hat{\beta}^{\mathsf{ols}}$.

where regularization is a critical issue may be an interesting area for future research.

2.4 Scaled Least Squares Estimator

Motivated by the results in the previous section, we design a computationally efficient algorithm that approximates the stochastic optimization problem Equation (2.1) that is as simple as solving the least squares problem; it is described in Algorithm 1. The algorithm has two basic steps. First, we estimate the OLS coefficients, and then in the second step we estimate the proportionality constant via a simple root-finding algorithm.

There are numerous fast optimization methods to solve the least squares problem, and even a superficial review of these could go beyond the page limits of this chapter. We emphasize that this step (finding the OLS estimator) does not have to be iterative and it is the main computational cost of the proposed algorithm. We suggest using a subsampling based estimator for β^{ols}, where we only use a subset of the observations to estimate the covariance matrix. Let $S \subset [n]$ be a random subsample and denote by $\mathbf{X}_S$ the sub-matrix formed by the rows of $\mathbf{X}$ in S. Then the subsampled OLS estimator is given as $\hat{\beta}^{\mathsf{ols}} = \left(\frac{1}{|S|}\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\frac{1}{n}\mathbf{X}^T y$. Properties of subsampling and sketching based estimators have been well-studied [Ver10, DLFU13, EM15, PW15, RKM16]. For sub-Gaussian covariates, it suffices to use a subsample size of $\mathcal{O}\left(p\log(p)\right)$ [Ver10]. Hence, this step requires a single time computational cost of $\mathcal{O}\left(|S|p^2 + p^3 + np\right) \approx \mathcal{O}\left(p\max\{p^2\log(p), n\}\right)$. For other approaches, we refer reader to [RT08, DMMS11, DLFU13, EM15] and the references therein.

The second step of Algorithm 1 involves solving a simple root-finding problem. As with

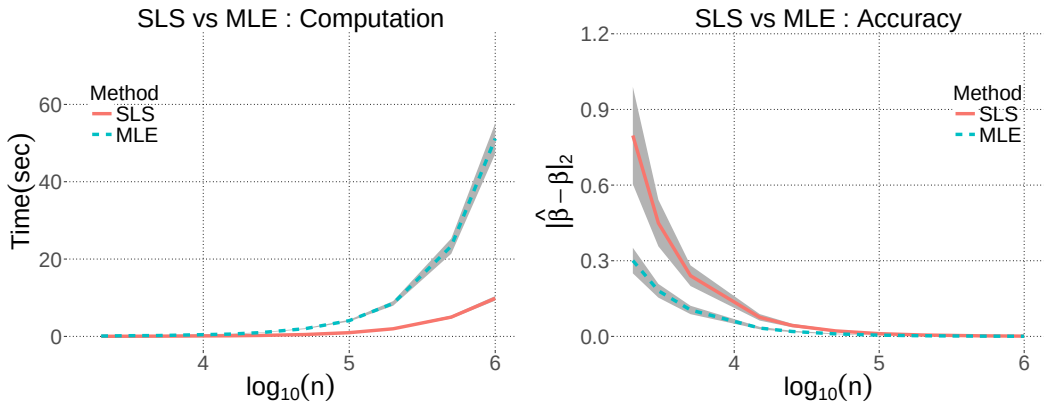

Figure 2.1: Logistic regression with iid standard Gaussian design. The left plot shows the computational cost (time) for finding the MLE and SLS as n grows and $p = 200$. The right plot depicts the accuracy of the estimators. In the regime where the MLE is expensive to compute, the SLS is found much more rapidly and has the same accuracy. `R`'s built-in functions are used to find the MLE.

the first step of the algorithm, there are numerous methods available for completing this task. Newton root-finding method with quadratic convergence or Halley's method with cubic convergence may be appropriate choices. We highlight that this step costs only $\mathcal{O}(n)$ per-iteration and that we can attain fast convergence rates of higher order algorithms. The resulting per-iteration cost is cheaper than other commonly used batch algorithms by at least a factor of $\mathcal{O}(p)$ — indeed, the cost of computing the gradient is $\mathcal{O}(np)$. For simplicity, we use Newton root-finding method.

Correct initialization of the scaling constant c depends on the optimization problem. For example, in the case of GLM problems, assuming that the GLM is a good approximation to the true conditional distribution, by the law of total variance and basic properties of GLMs, we have

$$\mathrm{Var}(y_i) = \mathbb{E}[\mathrm{Var}(y_i \mid x_i)] + \mathrm{Var}(\mathbb{E}[y_i \mid x_i]) \approx c_\Psi^{-1} + \mathrm{Var}\big(\Psi^{(1)}(\langle x_i, \beta\rangle)\big). \tag{2.11}$$

It follows that the initialization $c = 2/\mathrm{Var}(y_i)$ is reasonable as long as $c_\Psi^{-1} \approx \mathbb{E}[\mathrm{Var}(y_i \mid x_i)]$ is not much smaller than $\mathrm{Var}\big(\Psi^{(1)}(\langle x_i, \beta\rangle)\big)$. Our experiments show that SLS is very robust to initialization.

In Figure 2.1, we compare the performance of our SLS estimator to that of the MLE in a GLM optimization problem, when both are used to analyze synthetic data generated from a logistic regression model under general Gaussian design with randomly generated

covariance matrix. The left plot shows the computational cost of obtaining both estimators as n increases for fixed p. The right plot shows the accuracy of the estimators. In the regime $n \gg p \gg 1$ — where the MLE is hard to compute — the MLE and the SLS achieve the same accuracy, yet SLS has significantly smaller computation time. We refer the reader to Section 2.5 for theoretical results characterizing the finite sample behavior of the SLS.

2.5 Theoretical Results

In this section, we use the zero-bias transformations [GR97] to generalize the equivalence relation given in the previous section to the settings where the covariates are non-Gaussian.

Definition 1. *Let z be a random variable with mean 0 and variance σ^2. Then, there exists a random variable z^* that satisfies $\mathbb{E}\left[zf(z)\right] = \sigma^2\mathbb{E}[f^{(1)}(z^*)]$, for all differentiable functions f. The distribution of z^* is said to be the z-zero-bias distribution.*

The existence of z^* in Definition 1 is a consequence of Riesz representation theorem [GR97]. The normal distribution is the unique distribution whose zero-bias transformation is itself (i.e. the normal distribution is a fixed point of the operation mapping the distribution of z to that of z^* – which is basically Stein's lemma).

To provide some intuition behind the usefulness of the zero-bias transformation, we refer back to the proof of Proposition 2.3.1. For simplicity, assume that the covariate vector x_i has iid entries with mean 0, and variance 1. Then the zero-bias transformation applied to the j-th normal equation in Equation (2.7) yields

$$\underbrace{\mathbb{E}\left[y_i x_{ij}\right] = \mathbb{E}\left[x_{ij}\Psi^{(1)}\left(x_{ij}\beta_j + \Sigma_{k\neq j}x_{ik}\beta_k\right)\right]}_{j\text{-th normal equation}} = \underbrace{\beta_j\mathbb{E}\left[\Psi^{(2)}\left(x^*_{ij}\beta_j + \Sigma_{k\neq j}x_{ik}\beta_{ik}\right)\right]}_{\text{Zero-bias transformation}}. \quad (2.12)$$

The distribution of x^*_{ij} is the x_{ij}-zero-bias distribution and is entirely determined by the distribution of x_{ij}; general properties of x^*_{ij} can be found, for example, in [CGS10]. If β is well spread, it turns out that taken together, with $j = 1, \ldots, p$, the far right-hand side in Equation (2.12) behaves similar to the right side of Equation (2.8), with $\mathbf{\Sigma} = \mathbf{I}$; that is, the behavior is similar to the Gaussian case, where the proportionality relationship given in Proposition 2.3.1 holds. This argument leads to an approximate proportionality relationship for problems with non-Gaussian predictors, which, when carried out rigorously, yields the following result.

Theorem 2.5.1. *Suppose that the whitened covariates $w_i = \mathbf{\Sigma}^{-1/2}x_i$ are independent with mean 0, covariance $\mathbf{I}$, and have sub-Gaussian norm bounded by κ. Furthermore, w_i's have constant first and second conditional moments, i.e., $\forall j \in [p]$ and $\tilde{\beta} = \mathbf{\Sigma}^{1/2}\beta^{\mathsf{pop}}$, $\mathbb{E}[w_{ij}|\Sigma_{k\neq j}\tilde{\beta}_k w_{ik}]$ and $\mathbb{E}[w_{ij}^2|\Sigma_{k\neq j}\tilde{\beta}_k w_{ik}]$ are constant. Let $\|\beta^{\mathsf{pop}}\|_2 = \tau$ and assume β^{pop} is r-well-spread in the sense that $\tau/\|\beta^{\mathsf{pop}}\|_\infty = r\sqrt{p}$ for some $r \in (0,1]$, and the function $\Psi^{(2)}$ is Lipschitz continuous with constant k. Then, for $c_\Psi = 1/\mathbb{E}\left[\Psi^{(2)}(\langle x_i, \beta^{\mathsf{pop}}\rangle)\right]$, and $\rho = \rho_\infty(\mathbf{\Sigma}^{1/2})$ denoting the condition number of $\mathbf{\Sigma}^{1/2}$, we have*

$$\left\| \frac{1}{c_\Psi} \times \beta^{\mathsf{pop}} - \beta^{\mathsf{ols}} \right\|_\infty \leq \frac{\eta}{p}, \quad \textit{where} \quad \eta = 8k\kappa^3\rho\|\mathbf{\Sigma}^{1/2}\|_\infty(\tau/r)^2. \tag{2.13}$$

Theorem 2.5.1 is proved in the Appendix. It implies that the population parameters β^{ols} and β^{pop} are approximately equivalent up to a scaling factor, with an error bound of $\mathcal{O}(1/p)$. For the analysis above, we followed the standard convention in statistics by assuming that the covariates have norm of order p, i.e. $\mathbb{E}\left[\|x\|_2^2\right] = \mathcal{O}(p)$, and the coefficient vector has norm of order 1, i.e. $\tau^2 = \mathcal{O}(1)$ [DL91, HL93, HJS01]. On the other hand, several works in machine learning literature assume that $\mathbb{E}\left[\|x\|_2^2\right] = \mathcal{O}(1)$ and that $\tau^2 = \mathcal{O}(p)$ [KS09, KKSK11]. We note that both settings are theoretically equivalent and the objective is to make $\mathbb{E}[\langle x, \beta^{\mathsf{pop}}\rangle] = \mathcal{O}(p)$. The assumption that β^{pop} is well-spread can be relaxed with minor modifications. For example, if we have a sparse coefficient vector, where $\mathrm{supp}(\beta^{\mathsf{pop}}) = \{j;\ \beta_j^{\mathsf{pop}} \neq 0\}$ is the support set of β^{pop}, then Theorem 2.5.1 holds with p replaced by the size of the support set.

The assumptions on the conditional moments are the relaxed versions of assumptions that are commonly encountered in dimension reduction techniques. For example, sliced inverse regression methods assume that the first conditional moment $\mathbb{E}\left[x|\langle x, \beta\rangle\right]$ is linear in x for all β [LD89, Li91], which is satisfied by elliptically distributed random vectors. An important case that is not covered by these methods is the independent coordinate case, i.e., when the whitened covariates have independent, but not necessarily identical entries. It is straightforward to observe that this case satisfies the assumptions of Theorem 2.5.1. We refer reader to [LD09], for a good review of dimension reduction techniques and their corresponding assumptions. We also highlight that our moment assumptions can be relaxed further, at the expense of introducing some additional complexity into the results.

An interesting consequence of Theorem 2.5.1 and the remarks following the theorem is that whenever an entry of β^{pop} is zero, the corresponding entry of β^{ols} has to be small, and

conversely. For $\lambda \geq 0$, define the lasso coefficients

$$\beta_{\lambda}^{\mathsf{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\mathbb{E}\left[(y_i - \langle x_i, \beta \rangle)^2\right] + \lambda \left\|\beta\right\|_1 . \tag{2.14}$$

Corollary 2.5.2. *For any $\lambda \geq \eta/|\mathrm{supp}(\beta^{\mathsf{pop}})|$, if $\mathbb{E}\left[x_i\right] = 0$ and $\mathbb{E}\left[x_i x_i^T\right] = \mathbf{I}$, we have $\mathrm{supp}(\beta^{\mathsf{lasso}}) \subset \mathrm{supp}(\beta^{\mathsf{pop}})$. Further, if λ and β^{pop} also satisfy that $\forall j \in \mathrm{supp}(\beta^{\mathsf{pop}})$, $|\beta_j^{\mathsf{pop}}| > c_{\Psi}\left(\lambda + \eta/|\mathrm{supp}(\beta^{\mathsf{pop}})|\right)$, then we have $\mathrm{supp}(\beta^{\mathsf{lasso}}) = \mathrm{supp}(\beta^{\mathsf{pop}})$.*

So far in this section, we have only discussed properties of the population parameters, such as β^{pop} and β^{ols}. In the remainder of this section, we turn our attention to results for the estimators that are the main focus of this chapter; these results ultimately build on our earlier results, i.e. Theorem 2.5.1.

In order to precisely describe the performance of $\hat{\beta}^{\mathsf{sls}}$, we first need bounds on the OLS estimator. The OLS estimator has been studied extensively in the literature; however, for our purposes, we find it convenient to derive a new bound on its accuracy. While we have not seen this exact bound elsewhere, it is very similar to Theorem 5 of [DLFU13].

Proposition 2.5.3. *Assume that $\mathbb{E}\left[x_i\right] = 0$, $\mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}$, and that $\mathbf{\Sigma}^{-1/2} x_i$ and y_i are sub-Gaussian with norms κ and γ, respectively. For λ_{min} denoting the smallest eigenvalue of $\mathbf{\Sigma}$, and $|S| > \eta p$,*

$$\left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_2 \leq \eta \lambda_{min}^{-1/2} \sqrt{\frac{p}{|S|}}, \tag{2.15}$$

with probability at least $1 - 3e^{-p}$, where η depends only on γ and κ.

Proposition 2.5.3 is proved in the Supplementary Material. Our main result on the performance of $\hat{\beta}^{\mathsf{sls}}$ is given next.

Theorem 2.5.4. *Let the assumptions of Theorem 2.5.1 and Proposition 2.5.3 hold with $\mathbb{E}[\|\mathbf{\Sigma}^{-1/2} x\|_2] = \tilde{\mu}\sqrt{p}$. Further assume that the function $f(z) = z\mathbb{E}\left[\Psi^{(2)}(\langle x, \beta^{\mathsf{ols}} \rangle z)\right]$ satisfies $f(\bar{c}) > 1 + \bar{\delta}\sqrt{p}$ for some $\bar{c}$ and $\bar{\delta}$ such that the derivative of f in the interval $[0, \bar{c}]$ does not change sign, i.e., its absolute value is lower bounded by $\upsilon > 0$. Then, for n and $|S|$ sufficiently large, with probability at least $1 - 5e^{-p}$, we have*

$$\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_{\infty} \leq \eta_1 \frac{1}{p} + \eta_2 \sqrt{\frac{p}{\min\left\{n/\log(n), |S|\right\}}}, \tag{2.16}$$

where the constants η_1 and η_2 are defined by

$$\eta_1 = \eta k \bar{c} \kappa^3 \rho \|\mathbf{\Sigma}^{1/2}\|_\infty (\tau/r)^2 \tag{2.17}$$

$$\eta_2 = \eta \bar{c} \lambda_{min}^{-1/2} \left(1 + \upsilon^{-1} \lambda_{min}^{1/2} \|\beta^{\mathsf{ols}}\|_\infty \max\left\{(b + k/\tilde{\mu}), k\bar{c}\kappa\right\}\right), \tag{2.18}$$

and $\eta > 0$ is a constant depending on κ and γ.

Note that the convergence rate of the upper bound in Equation (2.16) depends on the sum of the two terms, both of which are functions of the data dimensions n and p. The first term on the right in Equation (2.16) comes from Theorem 2.5.1, which bounds the discrepancy between $c_\Psi \times \beta^{\mathsf{ols}}$ and β^{pop}. This term is small when p is large, and it does not depend on the number of observations n.

The second term in the upper bound Equation (2.16) comes from estimating β^{ols} and c_Ψ. This term is increasing in p, which reflects the fact that estimating β^{pop} is more challenging when p is large. As expected, this term is decreasing in n and $|S|$, i.e. larger sample size yields better estimates. When the full OLS solution is used ($|S| = n$), the second term becomes $\mathcal{O}(\sqrt{p\log(n)/n})$, which suggests that $n/\log(n)$ should be at least of order p for good performance. Also, note that there is a theoretical threshold for the subsampling size $|S|$, namely $\mathcal{O}(n/\log(n))$, beyond which further subsampling provides no improvement. This suggests that the subsampling size should be smaller than $\mathcal{O}(n/\log(n))$.

2.6 Converting One GLM to Another

In this section, we describe an efficient algorithm to transform a generalized linear model to another. It is often the case that a practitioner would like to change the loss function (equivalently the model) he/she uses based on its performance. When the dataset is large, training a new model from the scratch is computationally inefficient and will be time consuming. In the following, we will use the proportionality relation to transition between different loss functions.

Assume that a practitioner fitted a GLM using the loss function (or cumulant generating function) Ψ_1, but he/she would like to train a new model using the loss function Ψ_2. Instead of maximizing the log-likelihood based on Ψ_2, one can exploit the proportionality relation and obtain the coefficients for the new GLM problem. Denote by β_1^{pop} and β_2^{pop} the GLM

Algorithm 2 Conversion from one GLM to another

Input: Data $(y_i, x_i)_{i=1}^n$, and $\hat{\beta}_1^{\text{glm}}$

Step 1. Compute $\hat{y} = \mathbf{X}\hat{\beta}_1^{\text{glm}}$, and $\kappa = \frac{1}{n}\sum_{i=1}^n \Psi_1^{(2)}(\hat{y}_i)$.

Step 2. Solve the following equation for $\rho \in \mathbb{R}$: $\kappa = \frac{\rho}{n}\sum_{i=1}^n \Psi_2^{(2)}(\hat{y}_i\rho)$

Use Newton root-finding method:

Initialize $\rho = 1$;

Repeat until convergence:

$$\rho \leftarrow \rho - \frac{\rho\,\frac{1}{n}\sum_{i=1}^n \Psi_2^{(2)}(\rho\,\hat{y}_i) - \kappa}{\frac{1}{n}\sum_{i=1}^n \left\{\Psi_2^{(2)}(\rho\,\hat{y}_i) + \rho\,\hat{y}_i\Psi_2^{(3)}(\rho\,\hat{y}_i)\right\}}.$$

Output: $\hat{\beta}_2^{\text{glm}} = \rho \times \hat{\beta}_1^{\text{glm}}$.

coefficients corresponding to the loss functions Ψ_1 and Ψ_2, respectively. We have

$$\frac{1}{c_{\Psi_1}}\beta_1^{\mathsf{pop}} = \frac{1}{c_{\Psi_2}}\beta_2^{\mathsf{pop}} = \beta^{\mathsf{ols}},$$

that is, both coefficients are proportional to the OLS coefficients which does not depend on the loss function. Therefore, these coefficients β_1^{pop} and β_2^{pop} are also proportional to each other and we can write

$$\beta_2^{\mathsf{pop}} = \frac{c_{\Psi_2}}{c_{\Psi_1}}\,\beta_1^{\mathsf{pop}} \coloneqq \rho\,\beta_1^{\mathsf{pop}}, \tag{2.19}$$

where the proportionality constant between two GLM types turns out to be the ratio between c_{Ψ_1} and c_{Ψ_2}, i.e. $\rho = c_{\Psi_2}/c_{\Psi_1}$. Using the definition of c_{Ψ_2}, we write

$$\begin{aligned} 1 &= c_{\Psi_2}\,\mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_2^{\mathsf{pop}}\rangle)\right], \\ &= c_{\Psi_1}\rho\,\mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_1^{\mathsf{pop}}\rangle\rho)\right]. \end{aligned}$$

Dividing the both sides by c_{Ψ_1} and using the equality $1/c_{\Psi_1} = \mathbb{E}\left[\Psi_1^{(2)}(\langle x, \beta_1^{\mathsf{pop}}\rangle)\right]$, we obtain

$$\mathbb{E}\left[\Psi_1^{(2)}(\langle x, \beta_1^{\mathsf{pop}}\rangle)\right] = \rho\,\mathbb{E}\left[\Psi_2^{(2)}(\langle x, \beta_1^{\mathsf{pop}}\rangle\rho)\right].$$

The above equation only involves β_1^{pop} as the coefficients (which is already assumed to be known or fitted by the practitioner). Therefore, if we solve it for the ratio ρ, we can estimate

Name	Loss function: $\ell(y;q)$	Weight	Can. Link: $q(z)$
Log-loss	$-y\log(q)-(1-y)\log(1-q)$	$\frac{1}{q(1-q)}$	$\frac{1}{1+\exp(-z)}$
Boosting loss	$y(q^{-1}-1)^{1/2}+(1-y)(q^{-1}-1)^{-1/2}$	$\frac{1}{[q(1-q)]^{3/2}}$	$\frac{1}{2}+\frac{z/2}{2(z^2/4+1)^{1/2}}$
Square loss	$y(1-q)^2+(1-y)q^2$	1	$\frac{1+z}{2}$

Table 2.1: Common loss functions and their canonical links

β_2^{pop} by simply using the proportionality relation given in Equation (2.19).

The procedure described above is summarized as Algorithm 2. We emphasize that this procedure does not require the computation of the OLS estimator which was the main cost of SLS. The procedure only requires a per-iteration cost of $\mathcal{O}(n)$. In other words, conversion from one GLM type to another is much simpler than obtaining the GLM coefficients from the scratch.

2.7 Binary Classification with Proper Scoring Rules

In this section, we assume that for $i \in [n]$, the response is binary $y_i \in \{0,1\}$. The binary classification problem can be described by the following minimization of an empirical risk

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}\ell(y_i;q(\langle x_i,\beta\rangle)), \tag{2.20}$$

where ℓ and q are referred to as the loss and the link functions, respectively. There are various loss functions that are used in practice. Examples include log-loss, boosting loss, square loss etc (See Table 2.1). As before, we constrain our analysis on the canonical links. The concept of canonical links for binary classification is introduced by [BSS05], and it is quite similar to the generalized linear problems.

For any given loss function, we define the partial losses $\ell_k(\cdot)=\ell(y=k;\cdot)$ for $k\in\{0,1\}$. Since we have a binary response variable, we can write any loss in the following format

$$\begin{aligned}\ell(y;q) =& y\ell_1(q)+(1-y)\ell_0(q),\\ =& y\left(\ell_1(q)-\ell_0(q)\right)+\ell_0(q).\end{aligned}$$

The above formulation is of the form of a generalized linear problem. Before moving forward, we recall the concept of proper scoring in binary classification, which is sometimes referred to as Fisher consistency.

Definition 2 (Proper scoring rules)**.** *Assume that $y \sim$ Bernoulli(η). If the expected loss $\mathbb{E}[\ell(y,q)]$ is minimized by $q = \eta$ for all $\eta \in (0,1)$, we call the loss function a proper scoring rule.*

The following theorem by [Sch89] provides a methodology for constructing a loss function for the proper scoring rules.

Theorem 2.7.1 ([Sch89])**.** *Let $w(dt)$ be a positive measure on $(0,1)$ that is finite on interval $(\epsilon, 1-\epsilon)$ $\forall \epsilon > 0$. Then the following defines a proper scoring rule*

$$\ell_1(q) = \int_q^1 (1-t)w(dt), \text{ and } \ell_0(q) = \int_0^q tw(dt).$$

The measure $w(dt)$ uniquely defines the loss function (generally referred to as the weight function, since all losses can be written as weighted average of cost weighted misclassification error [BSS05, RW10]). Examples of weight functions is given in Table 2.1. The above theorem has many interesting interpretations; one that is most useful to us is that $\ell_0^{(1)}(q) = qw(q)$.

The notion of canonical links for proper scoring rules are introduced by [BSS05], which corresponds to the notion of matching loss [HKW99, RW10]. The derivation of canonical links stems from the Hessian of the above minimization, which remedies two potential problems: non-convexity and asymptotic variance inflation. It turns out that by setting $w(q)q^{(1)}$ as constant, one can remedy both problems [BSS05]. We will skip the derivation and, without loss of generality, assume that the canonical link-loss pair satisfies $w(q)q^{(1)} = 1$. Note that any loss function has a natural canonical link. The following Theorem summarizes this concept.

Theorem 2.7.2 ([BSS05])**.** *For proper scoring rules with $w > 0$, there exists a canonical link function which is unique up to addition and multiplication by constants. Conversely, any link function is canonical for a unique proper scoring rule.*

The canonical link for a given loss function can be explicitly derived from the equation $w(q)q^{(1)} = 1$. We have provided some examples in Table 2.1. Using the definition of

canonical link for proper scoring rules, we write the normal equations $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbb{E}\left[\ell(y, q(\langle x, \beta\rangle))\right] = 0$ as

$$\begin{aligned}\mathbb{E}\left[xq^{(1)}(\langle x,\beta\rangle)\ell_0^{(1)}(q(\langle x,\beta\rangle))\right] &= \mathbb{E}\left[yxq^{(1)}(\langle x,\beta\rangle)\left(\ell_0^{(1)}(q(\langle x,\beta\rangle)) - \ell_1^{(1)}(q(\langle x,\beta\rangle))\right)\right] \\ \mathbb{E}\left[xq(\langle x,\beta\rangle)\right] &= \mathbb{E}\left[yx\right] \\ \boldsymbol{\Sigma}\beta\mathbb{E}\left[q^{(1)}(\langle x,\beta\rangle)\right] &= \mathbb{E}\left[yx\right].\end{aligned}$$

The last equation provides us with the analog of the proportionality relation we observed in generalized linear problems. In this case, we observe that the proportionality constant becomes $1/\mathbb{E}\left[q^{(1)}(\langle x,\beta\rangle)\right]$. Therefore, our algorithm can be used to obtain a fast training procedure for the binary classification problems under canonical links as well.

2.8 Canonicalization of the Square Loss

In this section, we present a method to approximate the square loss minimization problem using a canonical form. Using this canonical approximation, we can use the techniques developed in the previous sections to gain computational benefits. Consider a minimization problem of the following form

$$\underset{\beta}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\langle x_i,\beta\rangle)]^2. \tag{2.21}$$

The above problem is commonly encountered in many machine learning tasks – specifically, in the context of neural networks, the function f is called the activation function. Here, we consider a toy example to demonstrate how our methodology can be useful in a minimization problem of the above form.

We first use Taylor series expansion around a point θ (which should be close to $\langle x,\beta\rangle$), in order to approximate the function $f(z)$ with a linear function around $f(\theta)$. We write

$$\begin{aligned}\min_{\beta}(y - f(\langle x,\beta\rangle))^2 &= \min_{\beta} f(\langle x,\beta\rangle)^2 - 2yf(\langle x,\beta\rangle) \\ &\approx \min_{\beta}\frac{f(\langle x,\beta\rangle)^2}{2f'(\theta)} - y\langle x,\beta\rangle.\end{aligned} \tag{2.22}$$

Then, we obtain

$$\Psi(z) = \frac{f(z)^2}{2f'(\theta)}, \tag{2.23}$$

and the proportionality relation given in previous sections would hold approximately. The above approximation will be accurate when the activation function is smooth around the user-specified point θ. We suggest to use $\theta = 0$ since when p is large and β is well-spread, the inner product $\langle x, \beta \rangle$ should be close to its expectation $\mathbb{E}[\langle x, \beta \rangle] = 0$. This method can be used to derive proportionality relations for GLMs with non-canonical links (conditional on the link being nice), and also may be of interest in non-convex optimization.

2.9 Experiments

This section contains the results of a variety of numerical studies, which show that the Scaled Least Squares estimator reaches the minimum achievable test error substantially faster than commonly used batch algorithms for finding the MLE. Both logistic and Poisson regression models (two types of GLMs) are utilized in our analyses, which are based on several synthetic and real datasets.

For the convenience of the reader, we briefly describe the optimization algorithms for the MLE that were used in the experiments. For a detailed discussion, we refer to Section 1.2, and the references there-in.

1. *Newton-Raphson (NR)* achieves locally quadratic convergence by scaling the gradient by the inverse of the Hessian evaluated at the current iterate. Computing the Hessian has a per-iteration cost of $\mathcal{O}\left(np^2\right)$, which makes it impractical for large-scale datasets.

2. *Newton-Stein (NS)* is a recently proposed second-order batch algorithm specifically designed for GLMs [Erd15, Erd16]. The algorithm uses Stein's lemma and subsampling to efficiently estimate the Hessian with a cost of $\mathcal{O}(np)$ per-iteration, achieving near quadratic rates.

3. *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* is the most popular and stable quasi-Newton method [Nes13]. At each iteration, the gradient is scaled by a matrix that is formed by accumulating information from previous iterations and gradient computations. The convergence is locally super-linear with a per-iteration cost of $\mathcal{O}(np)$.

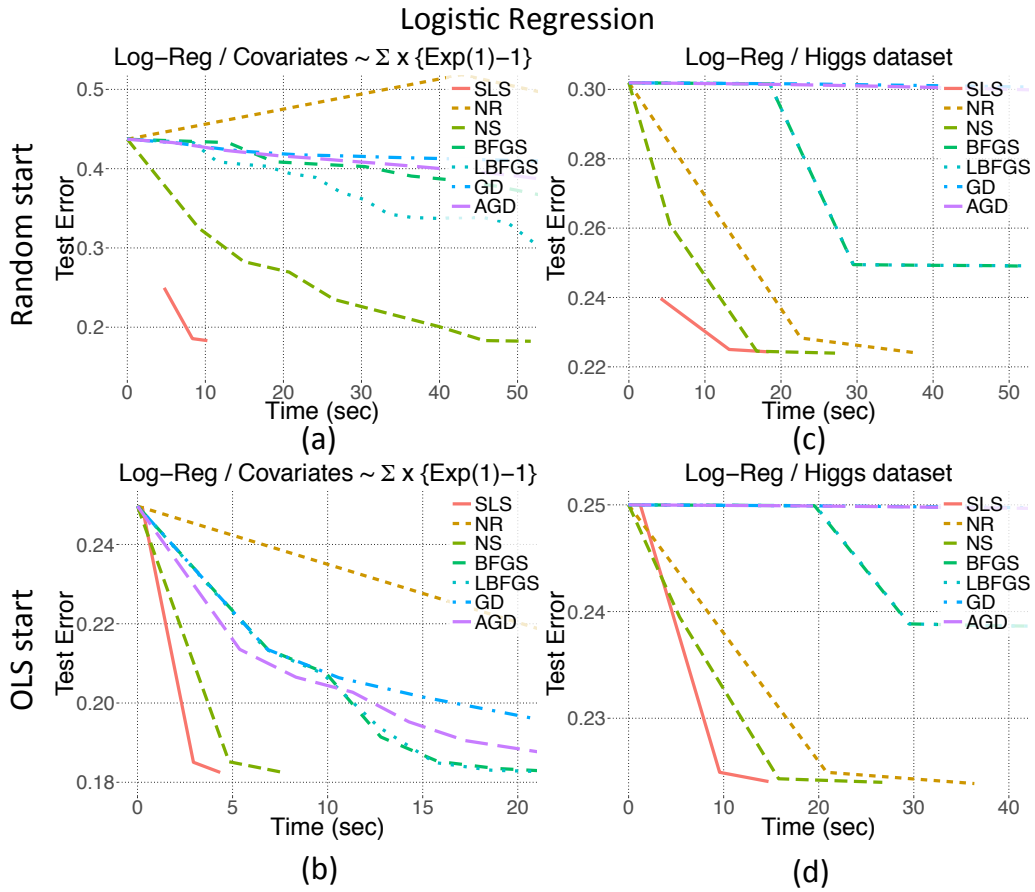

Figure 2.2: We compared the performance of SLS to that of MLE for the logistic regression problem on several datasets. MLE optimization is solved by various optimization algorithms. SLS is represented with red straight line. The details are provided in Table 2.2.

4. *Limited memory BFGS (LBFGS)* is a variant of BFGS, which uses only the recent iterates and gradients to approximate the Hessian, providing significant improvement in terms of memory usage. LBFGS has many variants; we use the formulation given in [Bis95].

5. *Gradient descent (GD)* takes a step in the opposite direction of the gradient, evaluated at the current iterate. Its performance strongly depends on the condition number of the design matrix. Under certain assumptions, the convergence is linear with $\mathcal{O}(np)$ per-iteration cost.

6. *Accelerated gradient descent (AGD)* is a modified version of gradient descent with

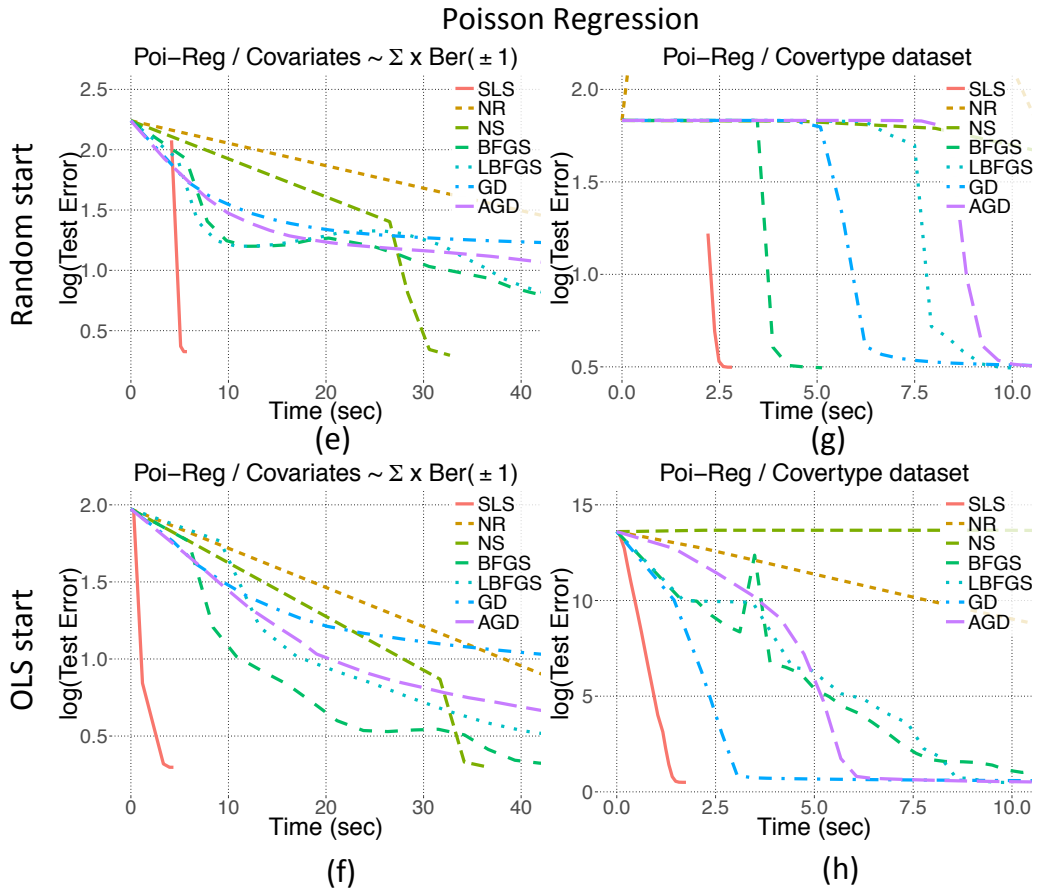

Figure 2.3: We compared the performance of SLS to that of MLE for the Poisson regression problem on several datasets. MLE optimization is solved by various optimization algorithms. SLS is represented with red straight line. The details are provided in Table 2.2.

an additional "momentum" term [Nes83]. Its per iteration cost is $\mathcal{O}(np)$ and its performance strongly depends on the smoothness of the objective function.

For all the algorithms for computing the MLE, the step size at each iteration is chosen via the backtracking line search [BV04].

Recall that the proposed Algorithm 1 is composed of two steps; the first finds an estimate of the OLS coefficients. This up-front computation is not needed for any of the MLE algorithms described above. On the other hand, each of the MLE algorithms requires some initial value for β, but no such initialization is needed to find the OLS estimator in Algorithm 1. This raises the question of how the MLE algorithms should be initialized, in

Model	Logistic regression				Poisson regression			
Dataset	$\mathbf{\Sigma}\times$\{Exp(1)-1\}		Higgs [BSW14]		$\mathbf{\Sigma}\times$Ber($\pm$1)		Covertype [BD99]	
Size	$n = 6.0 \times 10^5$, $p = 300$		$n = 1.1\times10^7$, $p = 29$		$n = 6.0\times10^5$, $p = 300$		$n = 5.8\times10^5$, $p = 53$	
Init	Rnd	Ols	Rnd	Ols	Rnd	Ols	Rnd	Ols
Plot	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Method	Time in seconds / number of iterations (to reach min test error)							
Sls	8.34/4	2.94/3	13.18/3	9.57/3	5.42/5	3.96/5	2.71/6	1.66/20
Nr	301.06/6	82.57/3	37.77/3	36.37/3	170.28/5	130.1/4	16.7/8	32.48/18
Ns	51.69/8	7.8/3	27.11/4	26.69/4	32.71/5	36.82/4	21.17/10	282.1/216
Bfgs	148.43/31	24.79/8	660.92/68	701.9/68	67.24/29	72.42/26	5.12/7	22.74/59
Lbfgs	125.33/39	24.61/8	6368.1/651	6946.1/670	224.6/106	357.1/88	10.01/14	10.05/17
Gd	669/138	134.91/25	100871/10101	141736/13808	1711/513	1364/374	14.35/25	33.58/87
Agd	218.1/61	35.97/12	2405.5/251	2879.69/277	103.3/51	102.74/40	11.28/15	11.95/25

Table 2.2: Details of the experiments shown in Figures 2.2 and 2.3.

order to compare them fairly with the proposed method. We consider two scenarios in our experiments: first, we use the OLS estimator computed for Algorithm 1 to initialize the MLE algorithms; second, we use a random initial value.

On each dataset, the main criterion for assessing the performance of the estimators is how rapidly the minimum test error is achieved. The test error is measured as the mean squared error of the estimated mean using the current parameters at each iteration on a test dataset, which is a randomly selected (and set-aside) 10% portion of the entire dataset. As noted previously, the MLE is more accurate for small n (see Figure 2.1). However, in the regime considered here ($n \gg p \gg 1$), the MLE and the SLS perform very similarly in terms of their error rates; for instance, on the Higgs dataset, the SLS and MLE have test error rates of 22.40% and 22.38%, respectively. For each dataset, the minimum achievable test error is set to be the maximum of the final test errors, where the maximum is taken over all of the estimation methods. Let $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ be two randomly generated covariance matrices. The datasets we analyzed were: (i) a synthetic dataset generated from a logistic regression model with iid {exponential(1)–1} predictors scaled by $\mathbf{\Sigma}^{(1)}$; (ii) the Higgs dataset (logistic regression) [BSW14]; (iii) a synthetic dataset generated from a Poisson regression model with iid binary($\pm$1) predictors scaled by $\mathbf{\Sigma}^{(2)}$; (iv) the Covertype dataset (Poisson regression) [BD99].

In all cases, the SLS outperformed the alternative algorithms for finding the MLE by a large margin, in terms of computation. Detailed results may be found in Figures 2.2 and 2.3, and Table 2.2. We provide additional experiments with different datasets in the Supplementary for this chapter.

2.10 Proof of Main Results

In this section, we provide the details and the proofs of our technical results. For convenience, we briefly state the following definitions.

Definition 3 (Sub-Gaussian)**.** *For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be* sub-Gaussian *if it satisfies*

$$\sup_{m \geq 1} m^{-1/2} \mathbb{E}\left[|x|^m\right]^{1/m} \leq \kappa.$$

Smallest such κ is the sub-Gaussian norm of x and it is denoted by $\|x\|_{\psi_2}$. Similarly, a random vector $y \in \mathbb{R}^p$ is a sub-Gaussian vector *if there exists a constant κ' such that*

$$\sup_{v \in S^{p-1}} \|\langle y, v \rangle\|_{\psi_2} \leq \kappa'.$$

Definition 4 (Sub-exponential)**.** *For a given constant κ, a random variable $x \in \mathbb{R}$ is called* sub-exponential *if it satisfies*

$$\sup_{m \geq 1} m^{-1} \mathbb{E}\left[|x|^m\right]^{1/m} \leq \kappa.$$

Smallest such κ is the sub-exponential norm of x and it is denoted by $\|x\|_{\psi_1}$. Similarly, a random vector $y \in \mathbb{R}^p$ is a sub-exponential vector *if there exists a constant κ' such that*

$$\sup_{v \in S^{p-1}} \|\langle y, v \rangle\|_{\psi_1} \leq \kappa'.$$

2.10.1 Proof of Theorem 2.5.1

Proof. For simplicity, we denote the whitened covariate by $w = \mathbf{\Sigma}^{-1/2} x$. Since w is sub-Gaussian with norm κ, its j-th entry w_j has bounded third moment. That is,

$$\begin{aligned} \kappa &= \sup_{\|u\|_2 = 1} \|\langle u, w \rangle\|_{\psi_2}, \\ &\geq \|w_j\|_{\psi_2} = \sup_{m \geq 1} m^{-1/2} \mathbb{E}\left[|w_j|^m\right]^{1/m}, \\ &\geq \frac{1}{\sqrt{3}} \mathbb{E}\left[|w_j|^3\right]^{1/3}, \end{aligned} \tag{2.24}$$

where in the first step, we used $u = e_j$, the j-th standard basis vector. Hence, we obtain a bound on the third moment, i.e,

$$\max_j \mathbb{E}\left[|w_j|^3\right] \leq 3^{3/2}\kappa^3. \tag{2.25}$$

Using the normal equations, we write

$$\begin{aligned}\mathbb{E}\left[yx\right] = \mathbb{E}\left[x\Psi^{(1)}(\langle x, \beta\rangle)\right] =& \boldsymbol{\Sigma}^{1/2}\mathbb{E}\left[w\Psi^{(1)}(\langle w, \boldsymbol{\Sigma}^{1/2}\beta\rangle)\right], \\ =& \boldsymbol{\Sigma}^{1/2}\mathbb{E}\left[w\Psi^{(1)}(\langle w, \tilde{\beta}\rangle)\right],\end{aligned} \tag{2.26}$$

where we defined $\tilde{\beta} = \boldsymbol{\Sigma}^{1/2}\beta$. By multiplying both sides with $\boldsymbol{\Sigma}^{-1}$, we obtain

$$\beta^{\mathsf{ols}} = \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\left[w\Psi^{(1)}(\langle w, \tilde{\beta}\rangle)\right]. \tag{2.27}$$

Now we define the partial sums $W_{-i} = \sum_{j\neq i}\tilde{\beta}_j w_j = \langle\tilde{\beta}, w\rangle - \tilde{\beta}_i w_i$. We will focus on the i-th entry of the above expectation given in Equation (2.27). Denoting the zero biased transformation of w_i conditioned on W_{-i} by w_i^*, we have

$$\begin{aligned}\mathbb{E}\left[w_i\Psi^{(1)}(\langle w, \tilde{\beta}\rangle)\right] =& \mathbb{E}\left[\mathbb{E}\left[w_i\Psi^{(1)}\left(\tilde{\beta}_i w_i + W_{-i}\right)|W_{-i}\right]\right], \\ =& \tilde{\beta}_i\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i w_i^* + W_{-i})\right], \\ =& \tilde{\beta}_i\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right],\end{aligned} \tag{2.28}$$

where in the second step, we used the assumption on conditional moments. Let $\mathbf{D}$ be a diagonal matrix with diagonal entries $\mathbf{D}_{ii} = \mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right]$. Using Equation (2.27) together with Equation (2.28), we obtain the equality

$$\begin{aligned}\beta^{\mathsf{ols}} =& \boldsymbol{\Sigma}^{-1/2}\mathbf{D}\tilde{\beta}, \\ =& \boldsymbol{\Sigma}^{-1/2}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\beta.\end{aligned} \tag{2.29}$$

Now, using the Lipschitz continuity assumption of the variance function, we have

$$\left|\mathbb{E}\left[\Psi^{(2)}(\tilde{\beta}_i(w_i^* - w_i) + \langle w, \tilde{\beta}\rangle)\right] - \mathbb{E}\left[\Psi^{(2)}(\langle w, \tilde{\beta}\rangle)\right]\right| \leq k|\tilde{\beta}_i|\mathbb{E}\left[|w_i^* - w_i|\right]. \tag{2.30}$$

In the following, we will use the properties of zero-biased transformations. Consider the

quantity

$$r = \sup \frac{\mathbb{E}\left[|w_i^* - w_i| \middle| W_{-i}\right]}{\mathbb{E}\left[|w_i|^3 \middle| W_{-i}\right]} \tag{2.31}$$

where w_i^* has w_i-zero biased distribution (conditioned on W_{-i}) and the supremum is taken with respect to all random variables with mean 0, standard deviation 1 and finite third moment, and w_i^* is achieving the minimal ℓ_1 coupling to w_i conditioned on W_{-i}. It is shown in [Gol07] that the above bound holds for $r = 1.5$ for the unconditional zero-bias transformations. Here, we take a similar approach to show that the same bound holds for the conditional case as well. By using the triangle inequality, we have

$$\begin{aligned}\mathbb{E}\left[|w_i^* - w_i| \middle| W_{-i}\right] \leq & \mathbb{E}\left[|w_i^*| \middle| W_{-i}\right] + \mathbb{E}\left[|w_i| \middle| W_{-i}\right] \\ \leq & \frac{1}{2}\mathbb{E}\left[|w_i|^3 \middle| W_{-i}\right] + \mathbb{E}\left[|w_i|^3 \middle| W_{-i}\right]^{1/3}.\end{aligned}$$

Since $\mathbb{E}\left[|w_i|^2 \middle| W_{-i}\right]$ is constant, it is equal to $\mathbb{E}\left[|w_i|^2\right] = 1$. This yields that the second term in the last line is upper bounded by $\mathbb{E}\left[|w_i|^3 \middle| W_{-i}\right]$. Consequently, by taking expectations over both hand sides we obtain that

$$\mathbb{E}\left[|w_i^* - w_i|\right] \leq 1.5\ \mathbb{E}\left[|w_i|^3\right].$$

Then the right hand side of Equation (2.30) can be upper bounded by

$$\begin{aligned}k|\tilde{\beta}_i|\mathbb{E}\left[|w_i^* - w_i|\right] \leq & rk \max_i \left\{|\tilde{\beta}_i|\mathbb{E}\left[|w_i|^3\right]\right\}, \\ \leq & 1.5k \left\|\mathbf{\Sigma}^{1/2}\beta\right\|_\infty 3^{3/2}\kappa^3, \\ \leq & 8k\kappa^3\|\mathbf{\Sigma}^{1/2}\beta\|_\infty,\end{aligned} \tag{2.32}$$

where in the second step we used the bound on the third moment given in Equation (2.25). The last inequality provides us with the following result,

$$\max_i \left|\mathbf{D}_{ii} - \frac{1}{c_\Psi}\right| \leq 8k\kappa^3\|\mathbf{\Sigma}^{1/2}\beta\|_\infty. \tag{2.33}$$

Finally, combining this with Equation (2.27) and Equation (2.29), we obtain

$$\begin{aligned}\left\|\beta^{\mathsf{ols}} - \frac{1}{c_\Psi}\beta\right\|_\infty &= \left\|\mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{\Sigma}^{1/2}\beta - \frac{1}{c_\Psi}\beta\right\|_\infty, \\ &= \left\|\mathbf{\Sigma}^{-1/2}\left(\mathbf{D} - \frac{1}{c_\Psi}\mathbf{I}\right)\mathbf{\Sigma}^{1/2}\beta\right\|_\infty, \\ &\leq \max_i \left|\mathbf{D}_{ii} - \frac{1}{c_\Psi}\right| \left\|\mathbf{\Sigma}^{1/2}\right\|_\infty \left\|\mathbf{\Sigma}^{-1/2}\right\|_\infty \|\beta\|_\infty^2, \\ &\leq 8k\kappa^3\rho(\mathbf{\Sigma}^{1/2})\|\mathbf{\Sigma}^{1/2}\|_\infty\frac{\tau^2}{r^2 p},\end{aligned} \tag{2.34}$$

where in the last step, we used the assumption that β is r-well-spread. □

2.10.2 Proof of Proposition 2.5.3

Proof. For convenience, we denote the whitened covariates with $w_i = \mathbf{\Sigma}^{-1/2}x_i$. We have $\mathbb{E}[w_i] = 0$, $\mathbb{E}\left[w_i w_i^T\right] = \mathbf{I}$, and $\|w_i\|_{\psi_2} \leq \kappa$. Also denote the subsampled covariance matrix with $\widehat{\mathbf{\Sigma}} = \frac{1}{|S|}\sum_{i\in S} x_i x_i^T$, and its whitened version as $\widetilde{\mathbf{\Sigma}} = \frac{1}{|S|}\sum_{i\in S} w_i w_i^T$. Further, define $\hat{\zeta} = \frac{1}{n}\sum_{i=1}^n w_i y_i$ and $\zeta = \mathbb{E}[wy]$. Then, we have

$$\hat{\beta}^{\mathsf{ols}} = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}^{1/2}\hat{\zeta} \quad \text{and} \quad \beta^{\mathsf{ols}} = \mathbf{\Sigma}^{-1/2}\zeta.$$

For now, we work on the event that $\widehat{\mathbf{\Sigma}}$ is invertible. We will see that this event holds with very high probability. We write

$$\begin{aligned}\left\|\mathbf{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2 &= \left\|\mathbf{\Sigma}^{1/2}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}^{1/2}\hat{\zeta} - \mathbf{\Sigma}^{-1/2}\zeta\right\|_2, \\ &= \left\|\widetilde{\mathbf{\Sigma}}^{-1}\left\{\hat{\zeta} - \zeta + \left(\mathbf{I} - \mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2}\right)\zeta\right\}\right\|_2, \\ &\leq \left\|\widetilde{\mathbf{\Sigma}}^{-1}\right\|_2\left\{\left\|\hat{\zeta} - \zeta\right\|_2 + \left\|\mathbf{I} - \widetilde{\mathbf{\Sigma}}\right\|_2\|\zeta\|_2\right\},\end{aligned} \tag{2.35}$$

where we used the triangle inequality and the properties of the operator norm.

For the first term on the right hand side of Equation (2.35), we write

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}^{-1}\right\|_2 &= \frac{1}{\lambda_{\min}(\widetilde{\mathbf{\Sigma}})}, \\ &\leq \frac{1}{1-\delta},\end{aligned}$$

where we assumed that such a $\delta > 0$ exists. In fact, when $\delta < 0.5$, we obtain a bound of 2 on the right hand side, which also justifies the invertibility assumption of $\widehat{\mathbf{\Sigma}}$. By Lemma A.1.3 and the following remark, we have with probability at least $1 - 2\exp\{-p\}$,

$$\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{I}\right\|_2 \leq c\sqrt{\frac{p}{|S|}},$$

where c is a constant depending only on κ. When $|S| > 4c^2 p$, we obtain

$$\left|\lambda_{\min}(\widetilde{\mathbf{\Sigma}}) - 1\right| \leq \left\|\widetilde{\mathbf{\Sigma}} - \mathbf{I}\right\|_2 \leq 0.5,$$

where the first inequality follows from the Lipschitz property of the eigenvalues.

Next, we bound the difference between $\hat{\zeta}$ and its expectation ζ. We write the bounds on the sub-exponential norm

$$\begin{aligned}
\|wy\|_{\psi_1} =& \sup_{\|v\|_2=1} \sup_{m\geq 1} m^{-1}\mathbb{E}\left[|\langle v, w\rangle y|^m\right]^{1/m}, \\
\leq& \sup_{\|v\|_2=1} \sup_{m\geq 1} m^{-1}\mathbb{E}\left[|\langle v, w\rangle|^{2m}\right]^{1/2m} \mathbb{E}\left[|y|^{2m}\right]^{1/2m}, \\
\leq& \sup_{\|v\|_2=1} \sup_{m\geq 1} m^{-1/2}\mathbb{E}\left[|\langle v, w\rangle|^{2m}\right]^{1/2m} \sup_{m\geq 1} m^{-1/2}\mathbb{E}\left[|y|^{2m}\right]^{1/2m}, \\
\leq& 2\|w\|_{\psi_2}\|y\|_{\psi_2} = 2\gamma\kappa.
\end{aligned} \tag{2.36}$$

Hence, we have $\max_i \|w_i y_i - \mathbb{E}[w_i y_i]\|_{\psi_1} \leq 4\gamma\kappa$. Further, let e_j denote the j-th standard basis, and notice that each entry of w is also sub-Gaussian with norm upper bounded by κ, i.e.,

$$\begin{aligned}
\kappa = \|w\|_{\psi_2} =& \sup_{\|u\|_2=1} \|\langle u, w\rangle\|_{\psi_2}, \\
&\geq \|\langle e_j, w\rangle\|_{\psi_2} = \|w_j\|_{\psi_2}.
\end{aligned} \tag{2.37}$$

Also, we can write

$$\begin{aligned} 2\gamma\kappa \geq \|wy\|_{\psi_1} &= \sup_{\|u\|_2=1} \sup_{m\geq 1} m^{-1}\mathbb{E}\left[|\langle u, w\rangle y|^m\right]^{1/m}, \\ &\geq \sup_{\|u\|_2=1} \mathbb{E}\left[|\langle u, w\rangle y|\right], \\ &\geq \sup_{\|u\|_2=1} \mathbb{E}\left[\langle u, w\rangle y\right], \\ &= \sup_{\|u\|_2=1} \langle u, \zeta\rangle = \|\zeta\|_2, \end{aligned} \tag{2.38}$$

where in the last step, we used the fact that dual norm of ℓ_2 norm is itself.

Next, we apply Lemma A.1.1 to $\hat{\zeta} - \zeta$, and obtain with probability at least $1 - \exp\{-p\}$

$$\left\|\hat{\zeta} - \zeta\right\|_2 \leq c\gamma\kappa\sqrt{\frac{p}{n}},$$

whenever $n > c^2 p$ for an absolute constant c.

Combining the above results in Equation (2.35), we obtain with probability at least $1 - 3\exp\{-p\}$

$$\left\|\mathbf{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2 \leq 2\left\{c_1\gamma\kappa\sqrt{\frac{p}{n}} + c_2\gamma\kappa\sqrt{\frac{p}{|S|}}\right\} \leq \eta\sqrt{\frac{p}{|S|}} \tag{2.39}$$

where η depends only on κ and γ, and $|S| > \eta p$. Finally, we write

$$\begin{aligned} \left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_2 &\leq \lambda_{\min}^{-1/2}\left\|\mathbf{\Sigma}^{1/2}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2, \\ &\leq \eta\lambda_{\min}^{-1/2}\sqrt{\frac{p}{|S|}}, \end{aligned}$$

with probability at least $1 - 3\exp\{-p\}$, whenever $|S| > \eta p$. □

2.10.3 Proof of Theorem 2.5.4

The following lemma – combined with the Proposition 2.5.3 – provides the necessary tools to prove Theorem 2.5.4.

Lemma 2.10.1. *For a given function $\Psi^{(2)}$ that is Lipschitz continuous with k, and uniformly bounded by b, we define the function $f : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ as*

$$f(c, \beta) = c \, \mathbb{E}\left[\Psi^{(2)}(\langle x, \beta \rangle c)\right],$$

and its empirical counterpart as

$$\hat{f}(c, \beta) = c \, \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle x_i, \beta \rangle c).$$

Assume that for some $\delta, \bar{c} > 0$, we have $f(\bar{c}, \beta^{\mathsf{ols}}) \geq 1 + \delta$. Then, $\exists c_\Psi > 0$ satisfying the equation

$$1 = f(c_\Psi, \beta^{\mathsf{ols}}).$$

Further, assume that for some $\tilde{\delta} > 0$, we have $\delta = \tilde{\delta}\sqrt{p}$, and n and $|S|$ sufficiently large, i.e.,

$$\min\left\{\frac{n}{\log(n)}, |S|\right\} > K^2/\tilde{\delta}^2$$

for $K = \eta\bar{c}\max\{b + \kappa/\tilde{\mu}, k\bar{c}\kappa\}$. Then, with probability $1 - 5\exp\{-p\}$, there exists a constant $\hat{c}_\Psi \in (0, \bar{c})$ satisfying the equation

$$1 = \hat{c}_\Psi \, \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle x_i, \hat{\beta}^{\mathsf{ols}} \rangle \hat{c}_\Psi).$$

Moreover, if the derivative of $z \to f(z, \beta^{\mathsf{ols}})$ is bounded below in absolute value (i.e. does not change sign) by $\upsilon > 0$ in the interval $z \in [0, \bar{c}]$, then with probability $1 - 5\exp\{-p\}$, we have

$$|\hat{c}_\Psi - c_\Psi| \leq C\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}},$$

where $C = K/\upsilon$.

Proof of Lemma 2.10.1. First statement is obvious. We notice that $f(c, \beta^{\mathsf{ols}})$ is a continuous function in its first argument with $f(0, \beta^{\mathsf{ols}}) = 0$ and $f(\bar{c}, \beta^{\mathsf{ols}}) \geq 1 + \delta$. Hence, there exists

$c_\Psi > 0$ such that $f(c_\Psi, \beta^{\mathsf{ols}}) = 1$. If there are many solutions to the above equation, we choose the one that is closest to zero. The condition on the derivative will guarantee the uniqueness of the solution.

Next, we will show the existence of $\hat{c}_\Psi$ using a uniform concentration given by Lemma A.1.2. Define the ellipsoid centered around β^{ols} with radius δ,

$$\mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}}) = \left\{\beta : \left\|\mathbf{\Sigma}^{1/2}(\beta - \beta^{\mathsf{ols}})\right\|_2 \leq \delta\right\},$$

and the event $\mathcal{E}$ that $\hat{\beta}^{\mathsf{ols}}$ falls into $\mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})$, i.e.,

$$\mathcal{E} = \left\{\hat{\beta}^{\mathsf{ols}} \in \mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})\right\}.$$

By Proposition 2.5.3 and the inequality given in Equation (2.39), whenever $|S| > \eta p \max\left\{1, \eta/\delta^2\right\}$, we obtain

$$\mathbb{P}\left(\mathcal{E}^C\right) \leq 3\exp\left\{-p\right\},$$

where $\mathcal{E}^C$ denotes the complement of the event $\mathcal{E}$, and η is a constant depending only on κ and γ. For any $c \in [0, \bar{c}]$, on the event $\mathcal{E}$, we have

$$\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| \leq \sup_{\beta \in \mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})} \left|\hat{f}(c, \beta) - f(c, \beta)\right|.$$

Hence, we obtain the following inequality

$$\begin{aligned}
&\mathbb{P}\left(\sup_{c\in[0,\bar{c}]} \left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(\sup_{c\in[0,\bar{c}]} \left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \hat{\beta}^{\mathsf{ols}})\right| > \epsilon; \mathcal{E}\right) + \mathbb{P}\left(\mathcal{E}^C\right), \\
&\leq \mathbb{P}\left(\sup_{c\in[0,\bar{c}]} \sup_{\beta \in \mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})} \left|\hat{f}(c, \beta) - f(c, \beta)\right| > \epsilon\right) + 3\exp\left\{-p\right\}.
\end{aligned}$$

In the following, we will use Lemma A.1.2 for the first term in the last line above. Denoting

by w, the whitened covariates, we have $\langle x, \beta\rangle = \langle w, \mathbf{\Sigma}^{1/2}\beta\rangle$. Therefore,

$$\begin{aligned}
&\sup_{c\in[0,\bar{c}]}\sup_{\beta\in\mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})}\left|\hat{f}(c,\beta)-f(c,\beta)\right| \\
&\leq \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{\beta\in\mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle w_i,\mathbf{\Sigma}^{1/2}\beta\rangle c)-\mathbb{E}\left[\Psi^{(2)}(\langle w,\mathbf{\Sigma}^{1/2}\beta\rangle c)\right]\right|.
\end{aligned}$$

Next, define the ball centered around $\tilde{\beta}^{\mathsf{ols}} = \mathbf{\Sigma}^{1/2}\beta^{\mathsf{ols}}$, with radius δ as $\mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}}) = \mathbf{\Sigma}^{1/2}\mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})$. We have $\beta \in \mathcal{B}^{\delta}_{\mathbf{\Sigma}}(\beta^{\mathsf{ols}})$ if and only if $\mathbf{\Sigma}^{1/2}\beta \in \mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}})$. Then, the right hand side of the above inequality can be written as

$$\begin{aligned}
&\bar{c}\sup_{c\in[0,\bar{c}]}\sup_{\beta\in\mathcal{B}_{\delta}(\tilde{\beta}^{\mathsf{ols}})}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle w_i,\beta\rangle c)-\mathbb{E}\left[\Psi^{(2)}(\langle w,\beta\rangle c)\right]\right|, \\
&=\bar{c}\sup_{\beta\in\mathcal{B}_{\bar{c}\delta}(\tilde{\beta}^{\mathsf{ols}})}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle w_i,\beta\rangle)-\mathbb{E}\left[\Psi^{(2)}(\langle w,\beta\rangle)\right]\right|.
\end{aligned}$$

Then, by Lemma A.1.2, we obtain

$$\mathbb{P}\left(\sup_{c\in[0,\bar{c}]}\left|\hat{f}(c,\hat{\beta}^{\mathsf{ols}})-f(c,\hat{\beta}^{\mathsf{ols}})\right|>c'\bar{c}(b+\kappa/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}}\right)\leq 5\exp\{-p\} \tag{2.40}$$

whenever $np > 51\max\left\{\chi,\chi^{-1}\right\}$ where $\chi = (b+\kappa/\tilde{\mu})^2/(c'\delta^2k^2\bar{c}^2\tilde{\mu}^2)$.

Also, by the Lipschitz condition for $\Psi^{(2)}$, we have for any $c\in[0,\bar{c}]$, and β_1,β_2,

$$\begin{aligned}
|f(c,\beta_1)-f(c,\beta_2)| \leq& kc^2\mathbb{E}\left[\left|\langle w,\mathbf{\Sigma}^{1/2}(\beta_1-\beta_2)\rangle\right|\right] \\
\leq& k\bar{c}^2\kappa\left\|\mathbf{\Sigma}^{1/2}(\beta_1-\beta_2)\right\|_2.
\end{aligned}$$

Applying the above bound for $\beta_1 = \hat{\beta}^{\mathsf{ols}}$ and $\beta_2 = \beta^{\mathsf{ols}}$, we obtain with probability $1-3\exp\{-p\}$

$$\left|f(c,\hat{\beta}^{\mathsf{ols}})-f(c,\beta^{\mathsf{ols}})\right|\leq\eta k\bar{c}^2\kappa\sqrt{\frac{p}{|S|}}, \tag{2.41}$$

where the last step follows from Proposition 2.5.3 and the inequality given in Equation (2.39).

Combining this with the previous bound, and taking into account that $\mu = \tilde{\mu}\sqrt{p}$, for

any $c \in [0, \bar{c}]$, with probability $1 - 5\exp\{-p\}$, we obtain

$$\begin{aligned}\left|\hat{f}(c, \hat{\beta}^{\mathsf{ols}}) - f(c, \beta^{\mathsf{ols}})\right| \leq & c'\bar{c}(b + \kappa/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}} + \eta k\bar{c}^2\kappa\sqrt{\frac{p}{|S|}} \\ \leq & K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}\end{aligned}$$

where $K = \eta\bar{c}\max\{b + \kappa/\tilde{\mu}, k\bar{c}\kappa\}$. Here, η depends only on κ and γ.

In particular, for $c = \bar{c}$ we observe that

$$\begin{aligned}\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) \geq & f(\bar{c}, \beta^{\mathsf{ols}}) - K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}} \\ \geq & 1 + \delta - K\sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}.\end{aligned}$$

Therefore, for sufficiently large n and $|S|$ satisfying

$$\min\left\{\frac{n}{\log(n)}, |S|\right\} > K^2/\tilde{\delta}^2$$

we obtain $\hat{f}(\bar{c}, \hat{\beta}^{\mathsf{ols}}) > 1$. Since this function is continuous and $\hat{f}(0, \hat{\beta}^{\mathsf{ols}}) = 0$, we obtain the existence of $\hat{c}_\Psi \in [0, \bar{c}]$ with probability at least $1 - 5\exp\{-p\}$.

Now, since $\hat{c}_\Psi$ and c_Ψ satisfy the equations $\hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) = f(c_\Psi, \beta^{\mathsf{ols}}) = 1$ (with high probability), by the inequality given in Equation (2.40), with probability at least $1 - 5\exp\{-p\}$, we obtain

$$\begin{aligned}\left|1 - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| = & \left|\hat{f}(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| \\ \leq & c'\bar{c}(b + \kappa/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}}.\end{aligned}$$

Also, by the same argument in Equation (2.41), and Proposition 2.5.3, we get

$$\begin{aligned}\left|f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - f(\hat{c}_\Psi, \beta^{\mathsf{ols}})\right| \leq & k\bar{c}^2\kappa\left\|\mathbf{\Sigma}(\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}})\right\|_2 \\ \leq & \eta k\bar{c}^2\kappa\sqrt{\frac{p}{|S|}}.\end{aligned}$$

Now, using the Taylor's series expansion of $c \to f(c, \beta^{\mathsf{ols}})$ around c_Ψ, and the assumption

on the derivative of f with respect to its first argument, we obtain

$$\begin{aligned}
\upsilon \left|\hat{c}_\Psi - c_\Psi\right| \leq & \left|f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(c_\Psi, \beta^{\mathsf{ols}})\right| \\
\leq & \left|f(\hat{c}_\Psi, \beta^{\mathsf{ols}}) - f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}})\right| + \left|f(\hat{c}_\Psi, \hat{\beta}^{\mathsf{ols}}) - 1\right| \\
\leq & \eta k \bar{c}^2 \kappa \sqrt{\frac{p}{|S|}} + c' \bar{c}(b + \kappa/\tilde{\mu}) \sqrt{\frac{p}{n/\log(n)}} \\
\leq & K \sqrt{\frac{p}{\min\left\{n/\log\left(n\right), |S|\right\}}}
\end{aligned}$$

with probability at least $1 - 5\exp\left\{-p\right\}$. Here, the constant K is the same as before

$$K = \eta \bar{c} \max\left\{b + \kappa/\tilde{\mu}, k\bar{c}\kappa\right\}.$$

□

Proof of Theorem 2.5.4. We have

$$\begin{aligned}
\left\|\hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}}\right\|_\infty = & \left\|\hat{c}_\Psi \hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_\infty, \\
\leq & \left\|c_\Psi \beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_\infty + \left\|\hat{c}_\Psi \hat{\beta}^{\mathsf{ols}} - c_\Psi \beta^{\mathsf{ols}}\right\|_\infty,
\end{aligned} \tag{2.42}$$

where we used the triangle inequality for the ℓ_∞ norm. The first term on the right hand side can be bounded using Theorem 2.5.1. We write

$$\left\|c_\Psi \beta^{\mathsf{ols}} - \beta^{\mathsf{pop}}\right\|_\infty \leq \eta_1 \frac{1}{p}, \tag{2.43}$$

for $\eta_1 = 8k\bar{c}\kappa^3 \rho(\mathbf{\Sigma}^{1/2}) \|\mathbf{\Sigma}^{1/2}\|_\infty (\tau/r)^2$.

For the second term, we write

$$\begin{aligned}
\left\|\hat{c}_\Psi \hat{\beta}^{\mathsf{ols}} - c_\Psi \beta^{\mathsf{ols}}\right\|_\infty = & \left\|\hat{c}_\Psi \hat{\beta}^{\mathsf{ols}} \pm \hat{c}_\Psi \beta^{\mathsf{ols}} - c_\Psi \beta^{\mathsf{ols}}\right\|_\infty, \\
\leq & \left\|\hat{c}_\Psi \hat{\beta}^{\mathsf{ols}} - \hat{c}_\Psi \beta^{\mathsf{ols}}\right\|_\infty + \left\|\hat{c}_\Psi \beta^{\mathsf{ols}} - c_\Psi \beta^{\mathsf{ols}}\right\|_\infty, \\
\leq & \left|\hat{c}_\Psi\right| \left\|\hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}}\right\|_\infty + \left|\hat{c}_\Psi - c_\Psi\right| \left\|\beta^{\mathsf{ols}}\right\|_\infty,
\end{aligned} \tag{2.44}$$

where the first step follows from triangle inequality. By Lemma 2.10.1, for sufficiently large n and $|S|$, with probability $1 - 5\exp\left\{-p\right\}$, the constant $\hat{c}_\Psi$ exists and it is in the interval

$(0, \bar{c}]$. By the same lemma, with probability $1 - 5\exp\{-p\}$, we have

$$|\hat{c}_\Psi - c_\Psi| \leq \eta_4 \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}, \tag{2.45}$$

where $\eta_4 = \eta' \upsilon^{-1} \bar{c} \max\{b + \kappa/\tilde{\mu}, k\bar{c}\kappa\}$, for some constant η' depending on the sub-Gaussian norms κ and γ.

Also, by the norm equivalence and Proposition 2.5.3, we have with probability $1 - 3\exp\{-p\}$

$$\left\| \hat{\beta}^{\mathsf{ols}} - \beta^{\mathsf{ols}} \right\|_\infty \leq \eta_3 \sqrt{\frac{p}{|S|}}, \tag{2.46}$$

for $\eta_3 = \eta'' \lambda_{\min}^{-1/2}$, where η'' is constant depending only on γ and κ.

Finally, combining all these inequalities with the last line of Equation (2.42), we have with probability $1 - 5\exp\{-p\}$,

$$\begin{aligned} \left\| \hat{\beta}^{\mathsf{sls}} - \beta^{\mathsf{pop}} \right\|_\infty &\leq \eta_1 \frac{1}{p} + \eta_3 \bar{c} \sqrt{\frac{p}{|S|}} + \eta_4 \left\| \beta^{\mathsf{ols}} \right\|_\infty \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}, \\ &\leq \eta_1 \frac{1}{p} + \left(\eta_3 \bar{c} + \eta_4 \left\| \beta^{\mathsf{ols}} \right\|_\infty \right) \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}, \\ &= \eta_1 \frac{1}{p} + \eta_2 \sqrt{\frac{p}{\min\{n/\log(n), |S|\}}}, \end{aligned} \tag{2.47}$$

where

$$\begin{aligned} \eta_1 &= 8k\bar{c}\kappa^3 \rho(\mathbf{\Sigma}^{1/2}) \|\mathbf{\Sigma}^{1/2}\|_\infty (\tau/r)^2 \\ \eta_2 &= \eta_3 \bar{c} + \eta_4 \left\| \beta^{\mathsf{ols}} \right\|_\infty, \\ &= \eta \bar{c} \lambda_{\min}^{-1/2} \left(1 + \upsilon^{-1} \lambda_{\min}^{1/2} \|\beta^{\mathsf{ols}}\|_\infty \max\{(b + k/\tilde{\mu}), k\bar{c}\kappa\} \right). \end{aligned} \tag{2.48}$$

□

2.10.4 Proof of Corollary 2.5.2

Proof. The normal equations for the lasso minimization yields

$$\mathbb{E}\left[xx^T\right] \beta_\lambda^{\mathsf{lasso}} - \beta^{\mathsf{ols}} + \lambda s = 0,$$

where $s \in \partial \left\|\beta_\lambda^{\mathsf{lasso}}\right\|_1$. It is well-known that under the orthogonal design where the covariates have i.i.d. entries, the above equation reduces to

$$\mathsf{soft}(\beta^{\mathsf{ols}}; \lambda) = \beta_\lambda^{\mathsf{lasso}},$$

where $\mathsf{soft}(\,\cdot\,;\lambda)$ denotes the soft thresholding operator at level λ. For any $\beta \in \mathbb{R}^p$, let $\mathrm{supp}(\beta)$ denote the support of β, i.e., the set $\{i \in [p] : \beta_i \neq 0\}$. We have

$$\begin{aligned}\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) &= \{i \in [p] : \beta_{\lambda,i}^{\mathsf{lasso}} \neq 0\},\\ &= \{i \in [p] : |\beta_i^{\mathsf{ols}}| > \lambda\}\end{aligned}$$

By Theorem 2.5.1, we have

$$|\beta_i^{\mathsf{ols}}| \leq \frac{1}{c_\Psi}|\beta_i^{\mathsf{pop}}| + \frac{\eta}{|\mathrm{supp}(\beta^{\mathsf{pop}})|},$$

which implies that

$$\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) \subset \left\{i \in [p] : \frac{1}{c_\Psi}|\beta_i^{\mathsf{pop}}| + \frac{\eta}{|\mathrm{supp}(\beta^{\mathsf{pop}})|} > \lambda\right\}.$$

Hence, whenever $\lambda > \eta/|\mathrm{supp}(\beta^{\mathsf{pop}})|$, we have

$$\mathrm{supp}(\beta_\lambda^{\mathsf{lasso}}) \subset \mathrm{supp}(\beta^{\mathsf{pop}}).$$

Further, we have by Theorem 2.5.1

$$\frac{1}{c_\Psi}|\beta_i^{\mathsf{pop}}| \leq |\beta_i^{\mathsf{ols}}| + \frac{\eta}{|\mathrm{supp}(\beta^{\mathsf{pop}})|}.$$

Hence, whenever $|\beta_i^{\mathsf{pop}}| > c_\Psi\left(\lambda + \eta/|\mathrm{supp}(\beta^{\mathsf{pop}})|\right)$, we get $|\beta_i^{\mathsf{ols}}| > \lambda$. If this condition is satisfied for any entry in the support of β^{pop}, the corresponding lasso coefficient will be non-zero. Therefore, we get

$$\mathrm{supp}(\beta^{\mathsf{pop}}) \subset \mathrm{supp}(\beta_\lambda^{\mathsf{lasso}})$$

under this assumption. Combining this with the previous result, we conclude the proof. □

2.11 Discussion

In this chapter, we showed that the true minimizer of a generalized linear problem and the OLS estimator are approximately proportional under the general random design setting. Using this relation, we proposed a computationally efficient algorithm for large-scale problems that achieves the same accuracy as the empirical risk minimizer by first estimating the OLS coefficients and then estimating the proportionality constant through iterations that can attain quadratic or cubic convergence rate, with only $\mathcal{O}(n)$ per-iteration cost.

We briefly mentioned that the proportionality between the coefficients holds even when there is regularization in Section 2.3.1. Further pursuing this idea may be interesting for large-scale problems where regularization is crucial. Another interesting line of research is to find similar proportionality relations between the parameters in other large-scale optimization problems such as support vector machines. Such relations may reduce the problem complexity significantly.

Chapter 3

Second Order Stein Approximations to Hessian

Contents of this chapter are based on the papers [Erd15, Erd16]. In this chapter, we propose an alternative way of constructing the curvature information by formulating it as an estimation problem and applying a *Stein-type lemma* to the population Hessian, which allows further improvements through subsampling and eigenvalue thresholding. The algorithm is called Newton-Stein method, and it enjoys fast convergence rates, resembling that of second order methods, with modest per-iteration cost. We provide its convergence analysis for the general case where the rows of the design matrix are samples from a sub-Gaussian distribution. We show that the convergence has two phases, a quadratic phase followed by a linear phase.

3.1 Introduction

In this chapter, we focus on how to solve the maximum likelihood problem efficiently in the GLM setting when the number of observations n is much larger than the dimension of the coefficient vector p, i.e., $n \gg p \gg 1$. GLM optimization task is typically expressed as a minimization problem where the objective function is the negative log-likelihood that is denoted by $f(\beta)$ where $\beta \in \mathbb{R}^p$ is the coefficient vector. Many optimization algorithms are available for such minimization problems [Bis95, BV04, Nes13]. However, only a few uses the special structure of GLMs. In this chapter, we consider updates that are specifically

designed for GLMs, which are of the from

$$\beta \leftarrow \beta - \gamma \mathbf{Q} \nabla_\beta f(\beta)\,, \tag{3.1}$$

where γ is the step size and $\mathbf{Q}$ is a scaling matrix which provides curvature information.

For the updates of the form Equation (3.1), the performance of the algorithm is mainly determined by the scaling matrix $\mathbf{Q}$. We have seen in Section 1.2 in detail that classical *Newton method* and *natural gradient descent* can be recovered by simply taking $\mathbf{Q}$ to be the inverse Hessian and the inverse Fisher's information at the current iterate, respectively [Ama98, Nes13]. Second order methods may achieve quadratic convergence rate, yet they suffer from excessive cost of computing the scaling matrix at every iteration. On the other hand, if we take $\mathbf{Q}$ to be the identity matrix, we recover the standard *gradient descent* which has a linear convergence rate. Although the convergence rate of gradient descent is considered slow compared to that of second order methods such as Newton method, modest per-iteration cost makes it practical for large-scale optimization.

Section 1.1.1 briefly discusses the GLM framework and its relevant properties. The rest of the chapter is organized as follows: Section 3.1.1 surveys the related work and Section 3.2 introduces the notations we use throughout the chapter. In Section 3.3, we introduce Newton-Stein method, develop its intuition, and discuss the computational aspects. Section 3.4 covers the theoretical results and in Section 3.4.4 we discuss how to choose the algorithm parameters. Section 3.5 provides the empirical results where we compare the proposed algorithm with several other methods on four data sets. Finally, in Section 3.7, we conclude with a brief discussion along with a few future research directions.

3.1.1 Related Work

There are numerous optimization techniques that can be used to find the maximum likelihood estimator in GLMs. For moderate values of n and p, the classical second order methods such as Newton method (also referred to as Newton-Raphson) are commonly used. In large-scale problems, data dimensionality is the main factor while determining the optimization method, which typically falls into one of two major categories: online and batch methods. Online methods use a gradient (or sub-gradient) of a single, randomly selected observation to update the current iterate [RM51]. Their per-iteration cost is independent of n, but

the convergence rate might be extremely slow. There are several extensions of the classical stochastic descent algorithms, providing significant improvement and improved stability [Bot10, DHS11, SLRB, KEO15].

On the other hand, batch algorithms enjoy faster convergence rates, though their per-iteration cost may be prohibitive. In particular, second order methods enjoy quadratic convergence, but constructing the Hessian matrix generally requires excessive amount of computation. To remedy this issue, most research is focused on designing an approximate and cost-efficient scaling matrix. This idea lies at the core of Quasi-Newton methods such as BFGS [Bis95, Nes13].

Another approach to construct an approximate Hessian makes use of subsampling techniques [Mar10, BCNN11, VP12, EM15, RKM16]. Many contemporary learning methods rely on subsampling as it is simple and it provides significant boost over the first order methods. Further improvements through conjugate gradient methods and Krylov sub-spaces are available. Subsampling can also be used to obtain an approximate solution, with certain large deviation guarantees [DLFU13].

There are many composite variants of the aforementioned methods, that mostly combine two or more techniques. Well-known composite algorithms are the combinations of subsampling and Quasi-Newton [SYG07, BHNS14], stochastic and deterministic gradient descent [FS12], natural gradient and Newton method [LRF10], natural gradient and low-rank approximation [LRMB08], subsampling and eigenvalue thresholding [EM15].

Lastly, algorithms that specialize on certain types of GLMs include coordinate descent methods for the penalized GLMs [FHT10], trust region Newton-type methods [LWK08], and approximation methods [EBD16b, EBD16a].

3.2 Preliminaries and Notation

In this chapter, we consider a generalized linear problem in the following empirical risk form

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\ \widehat{R}(\beta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\langle x_i,\beta\rangle\right) - y_i\langle x_i,\beta\rangle\,. \tag{3.2}$$

The standard approach in minimizing functions of the above form is to use iterative methods. It is straightforward to write that the gradient and the Hessian of $\widehat{R}(\beta)$,

$$\nabla_\beta \widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{(1)}(\langle x_i,\beta\rangle)x_i - y_i x_i\right], \tag{3.3}$$

$$\boldsymbol{\nabla}^2_\beta \widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)x_i x_i^T. \tag{3.4}$$

For a sequence of scaling matrices $\{\mathbf{Q}^t\}_{t>0} \in \mathbb{R}^{p\times p}$, the traditional approach is to use iterations of the form

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \gamma_t \mathbf{Q}^t \nabla_\beta \widehat{R}(\hat{\beta}^t)$$

where γ_t is the step size. Unlike Chapter 2, our focus in this chapter will be updates of the above form. We will propose a new scaling matrix based on Stein's lemma.

We let $[n] = \{1, 2, ..., n\}$ and denote by $|S|$, the size of a set S. The gradient and the Hessian of f with respect to β are denoted by $\nabla_\beta f$ and $\boldsymbol{\nabla}^2_\beta f$, respectively. The j-th derivative of a function $f(w)$ is denoted by $f^{(j)}(w)$. For a vector x and a symmetric matrix $\mathbf{X}$, $\|x\|_2$ and $\|\mathbf{X}\|_2$ denote the ℓ_2 and spectral norms of x and $\mathbf{X}$, respectively. $\|x\|_{\psi_2}$ denotes the sub-Gaussian norm, which will be defined later. S^{p-1} denotes the p-dimensional sphere. $\mathcal{P}_\mathcal{C}$ denotes the projections onto the set $\mathcal{C}$, and $B_p(R) \subset \mathbb{R}^p$ denotes the p-dimensional ball of radius R. For a random variable x and density f, $x \sim f$ means that the distribution of x follows the density f. Multivariate Gaussian density with mean $\mu \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is denoted as $\mathsf{N}_p(\mu, \boldsymbol{\Sigma})$. For random variables x, y, $d(x, y)$ and $\mathbf{D}(x, y)$ denote probability metrics (will be explicitly defined) measuring the distance between the distributions of x and y. $\mathcal{N}(\cdots)$ and T_ϵ denote the bracketing number and ϵ-net.

3.3 Newton-Stein Method

Classical Newton-Raphson (or simply Newton) method is the standard approach for training GLMs for moderately large date sets. However, its per-iteration cost makes it impractical for large-scale optimization. The main bottleneck is the computation of the Hessian matrix that requires $\mathcal{O}\left(()\, np^2\right)$ flops which is prohibitive when $n \gg p \gg 1$. Numerous methods

Algorithm 3 Newton-Stein Method

Input: $\hat{\beta}^0, |S|, \epsilon, \{\gamma_t\}_{t\geq 0}$.

1. Estimate the covariance using a random subsample $S \subset [n]$:
   $\widehat{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{i\in S} x_i x_i^T$.
2. **while** $\left\|\hat{\beta}^{t+1} - \hat{\beta}^t\right\|_2 > \epsilon$ **do**
   $\hat{\mu}_2(\hat{\beta}^t) = \frac{1}{n}\sum_{i=1}^n \Psi^{(2)}(\langle x_i, \hat{\beta}^t\rangle), \qquad \hat{\mu}_4(\hat{\beta}^t) = \frac{1}{n}\sum_{i=1}^n \Psi^{(4)}(\langle x_i, \hat{\beta}^t\rangle),$
   $$\mathbf{Q}^t = \frac{1}{\hat{\mu}_2(\hat{\beta}^t)}\left[\widehat{\mathbf{\Sigma}}_S^{-1} - \frac{\hat{\beta}^t[\hat{\beta}^t]^T}{\hat{\mu}_2(\hat{\beta}^t)/\hat{\mu}_4(\hat{\beta}^t) + \langle \widehat{\mathbf{\Sigma}}_S\hat{\beta}^t, \hat{\beta}^t\rangle}\right],$$
   $\hat{\beta}^{t+1} = \hat{\beta}^t - \gamma_t\, \mathbf{Q}^t \nabla_\beta l(\hat{\beta}^t)$,
   $t \leftarrow t+1$.
3. **end while**

Output: $\hat{\beta}^t$.

have been proposed to achieve the fast convergence rate of Newton method while keeping the per-iteration cost manageable. To this end, a popular approach is to construct a scaling matrix $\mathbf{Q}^t$, which approximates the inverse Hessian at every iteration t.

The task of constructing an approximate Hessian can be viewed as an estimation problem. Assuming that the rows of $\mathbf{X}$ are i.i.d. random vectors, the Hessian of the negative log-likelihood of GLMs with a cumulant generating function ϕ has the following sample average form

$$\left[\mathbf{Q}^t\right]^{-1} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T \Psi^{(2)}(\langle x_i, \beta\rangle) \approx \mathbb{E}[xx^T\Psi^{(2)}(\langle x, \beta\rangle)]\,.$$

We observe that $\left[\mathbf{Q}^t\right]^{-1}$ is just a sum of i.i.d. matrices. Hence, the true Hessian is nothing but a sample mean estimator to its expectation. Another natural estimator would be the subsampled Hessian method which is extensively studied by [Mar10, BCNN11, EM15, RKM16]. Therefore, our goal is to propose an estimator for the population level Hessian that is also computationally efficient. Since n is large, the proposed estimator will be close to the true Hessian.

We use the following Stein-type lemma to find a more efficient estimator to the expectation of the Hessian.

Lemma 3.3.1 (Stein-type lemma)**.** *Assume that* $x \sim \mathsf{N}_p(0, \mathbf{\Sigma})$ *and* $\beta \in \mathbb{R}^p$ *is a constant vector. Then for any function* $f : \mathbb{R} \to \mathbb{R}$ *that is twice* "weakly" *differentiable, we have*

$$\mathbb{E}\left[xx^T f(\langle x, \beta\rangle)\right] = \mathbb{E}\left[f(\langle x, \beta\rangle)\right]\mathbf{\Sigma} + \mathbb{E}\left[f^{(2)}(\langle x, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\,. \tag{3.5}$$

Proof. The proof will follow from integration by parts. Let $g(x|\mathbf{\Sigma})$ denote the density of a multivariate normal random variable x with mean 0 and covariance $\mathbf{\Sigma}$. We recall the basic identity $xg(x|\mathbf{\Sigma})dx = -\mathbf{\Sigma}dg(x|\mathbf{\Sigma})$ and write

$$\begin{aligned}
\mathbb{E}[xx^T f(\langle x, \beta\rangle)] =& \int xx^T f(\langle x, \beta\rangle)g(x)dx,\\
=&\mathbf{\Sigma}\left\{\int f(\langle x, \beta\rangle)g(x|\mathbf{\Sigma})dx + \int \beta x^T f^{(1)}(\langle x, \beta\rangle)g(x|\mathbf{\Sigma})dx\right\},\\
=&\mathbf{\Sigma}\left\{\mathbb{E}[f(\langle x, \beta\rangle)] + \int \beta\beta^T f^{(2)}(\langle x, \beta\rangle)g(x|\mathbf{\Sigma})dx\mathbf{\Sigma}\right\},\\
=&\mathbb{E}[f(\langle x, \beta\rangle)]\mathbf{\Sigma} + \mathbb{E}\left[f^{(2)}(\langle x, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}.
\end{aligned}$$

□

The right hand side of Equation (3.5) is a rank-1 update to the first term. Hence, its inverse can be computed with $\mathcal{O}(p^2)$ cost. Quantities that change at each iteration are the ones that depend on β, i.e.,

$$\mu_2(\beta) = \mathbb{E}[\Psi^{(2)}(\langle x, \beta\rangle)], \quad \text{and} \quad \mu_4(\beta) = \mathbb{E}[\Psi^{(4)}(\langle x, \beta\rangle)].$$

Note that $\mu_2(\beta)$ and $\mu_4(\beta)$ are scalar quantities and they can be estimated by their corresponding sample means $\hat{\mu}_2(\beta)$ and $\hat{\mu}_4(\beta)$ (explicitly defined at Step 2 of Algorithm 1) respectively, with only $\mathcal{O}(np)$ computation.

To complete the estimation task suggested by Equation (3.5), we need an estimator for the covariance matrix $\mathbf{\Sigma}$. A natural estimator is the sample mean where, we only use a subsample of the indices $S \subset [n]$ so that the cost is reduced to $\mathcal{O}(|S|p^2)$ from $\mathcal{O}(np^2)$. Subsampling based sample mean estimator is denoted by $\widehat{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{i\in S} x_i x_i^T$, which is widely used in large-scale problems [Ver10]. We highlight the fact that Lemma 3.3.1 replaces $\mathcal{O}(np^2)$ per-iteration cost of Newton method with a one-time cost of $\mathcal{O}(np^2)$. We further

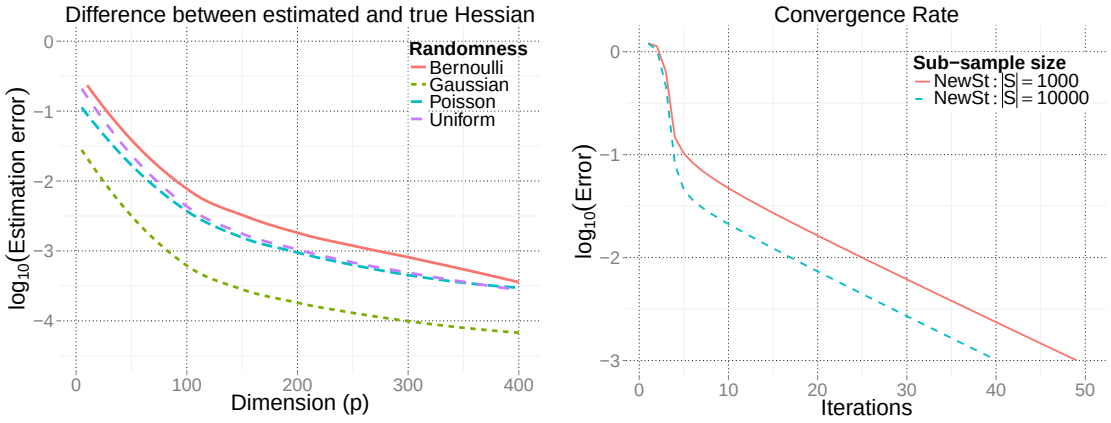

Figure 3.1: The left plot demonstrates the accuracy of proposed Hessian estimation over different distributions. Number of observations is set to be $n = \mathcal{O}(p\log(p))$. The right plot shows the phase transition in the convergence rate of Newton-Stein method (NewSt). Convergence starts with a quadratic rate and transitions into linear. Plots are obtained using *Covertype* data set.

use subsampling to reduce this one-time cost to $\mathcal{O}(|S|p^2)$, and obtain the following Hessian estimator at β

$$\underbrace{\left[\mathbf{Q}^t\right]^{-1}}_{\in \mathbb{R}^{p\times p}} = \underbrace{\hat{\mu}_2(\beta)}_{\in \mathbb{R}}\ \underbrace{\widehat{\mathbf{\Sigma}}_S}_{\in \mathbb{R}^{p\times p}} + \underbrace{\hat{\mu}_4(\beta)}_{\in \mathbb{R}}\ \overbrace{\underbrace{\widehat{\mathbf{\Sigma}}_S\beta\beta^T\widehat{\mathbf{\Sigma}}_S}_{\in \mathbb{R}^{p\times p}}}^{\text{rank-1 update}} \tag{3.6}$$

We emphasize that any covariance estimation method can be applied in the first step of the algorithm. There are various estimation techniques most of which rely on the concept of *shrinkage* [CCS10, DGJ13]. This is because, important curvature information is generally contained in the largest few spectral features [EM15]. In particular, for a given threshold r, we suggest to use the largest r eigenvalues of the subsampled covariance estimator $\widehat{\mathbf{\Sigma}}_S$, and setting rest of them to $(r+1)$-th eigenvalue. This operation helps denoising and provides additional computational benefits when inverting the covariance estimator [EM15].

Inverting the constructed Hessian estimator can make use of the low-rank structure. First, notice that the updates in Equation (3.6) are based on rank-1 matrix additions. Hence, we can simply apply Sherman–Morrison inversion formula to Equation (3.6) and obtain an explicit equation for the scaling matrix $\mathbf{Q}^t$ (Step 2 of Algorithm 1). This formulation would impose another inverse operation on the covariance estimator. We emphasize that

this operation is performed once. Therefore, instead of $\mathcal{O}(p^3)$ per-iteration cost of Newton method due to inversion, Newton-Stein method (NewSt) requires $\mathcal{O}(p^2)$ per-iteration and a one-time cost of $\mathcal{O}(p^3)$. Assuming that Newton-Stein and Newton methods converge in T_1 and T_2 iterations respectively, the overall complexity of Newton-Stein is

$$\mathcal{O}\left(npT_1 + p^2T_1 + (|S| + p)p^2\right) \approx \mathcal{O}\left(npT_1 + p^2T_1 + |S|p^2\right)$$

whereas that of Newton is $\mathcal{O}(np^2T_2 + p^3T_2)$. We show both empirically and theoretically that the quantities T_1 and T_2 are close to each other.

The convergence rate of Newton-Stein method has two phases. Convergence starts quadratically and transitions into linear rate when it gets close to the true minimizer. The phase transition behavior can be observed through the right plot in Figure 3.1. This is a consequence of the bound provided in Equation (1.26), which is the main result of our theorems on the local convergence (given in Section 3.4).

Even though Lemma 3.3.1 assumes that the covariates are multivariate Gaussian random vectors, in Section 3.4, the only assumption we make on the covariates is either bounded support or sub-Gaussianity, both of which cover a wide class of random variables including Bernoulli, elliptical distributions, bounded variables etc. The left plot of Figure 3.1 shows that the estimation is accurate for many distributions. This is a consequence of the fact that the proposed estimator in Equation 3.6 relies on the distribution of x only through inner products of the form $\langle x, v\rangle$, which in turn results in an approximate normal distribution due to the central limit theorem. To provide more intuition, we explain this through *zero-biased transformations* which is a general version of Stein's lemma for arbitrary distributions [GR97].

Definition 5. *Let z be a random variable with mean 0 and variance σ^2. Then, there exists a random variable z^* that satisfies $\mathbb{E}[zf(z)] = \sigma^2\mathbb{E}[f^{(1)}(z^*)]$, for all differentiable functions f. The distribution of z^* is said to be the z-zero-bias distribution.*

The normal distribution is the unique distribution whose zero-bias transformation is itself (i.e. the normal distribution is a fixed point of the operation mapping the distribution of z to that of z^*). The distribution of z^* is referred to as z-zero-bias distribution and is entirely determined by the distribution of z. Properties such as existence can be found, for example, in [CGS10].

To provide some intuition behind the usefulness of Lemma 3.3.1 even for arbitrary

distributions, we use zero-bias transformations. For simplicity, assume that the covariate vector x has i.i.d. entries from an arbitrary distribution with mean 0, and variance 1. Then the zero-bias transformation applied twice to the entry (i, j) of matrix $\mathbb{E}[xx^T f(\langle x, \beta\rangle)]$ yields

$$\mathbb{E}[x_i x_j f(\langle x, \beta\rangle)] = \begin{cases} \mathbb{E}[f(\beta_i x_i^* + \sum_{k\neq i} x_k\beta_k)] + \beta_i^2 \mathbb{E}\big[f^{(2)}(\beta_i x_i^{**} + \sum_{k\neq i} x_k\beta_k)\big] & \text{if} \quad i = j, \\ \beta_i\beta_j \mathbb{E}\big[f^{(2)}(\beta_i x_i^* + \beta_j x_j^* + \sum_{k\neq i,j} x_k\beta_k)\big] & \text{if} \quad i \neq j, \end{cases}$$

where x_i^* and x_i^{**} have x_i-zero-bias and x_i^*-zero-bias distributions, respectively. For each entry (i, j) at most two summands of $\langle x, \beta\rangle = \Sigma_k x_k\beta_k$ change their distributions. Therefore, if β is well spread and p is sufficiently large, the sums inside the expectations will behave similar to the inner product $\langle x, \beta\rangle$. Correspondingly, the above equations will be close to their Gaussian counterpart as given in Equation (3.5).

3.4 Theoretical Results

We start by introducing the terms that will appear in the theorems. Then we will provide two technical results on bounded and sub-Gaussian covariates. The proofs of the theorems are technical and provided in Appendix.

3.4.1 Preliminaries

Hessian estimation described in the previous section relies on a Gaussian approximation. For theoretical purposes, we use the following probability metric to quantify the gap between the distribution of x_i's and that of a normal vector.

Definition 6. *Given a family of functions $\mathcal{H}$, and random vectors $x, y \in \mathbb{R}^p$, for $\mathcal{H}$ and any $h \in \mathcal{H}$, define*

$$d_{\mathcal{H}}(x, y) = \sup_{h\in\mathcal{H}} d_h(x, y) \qquad \textit{where} \qquad d_h(x, y) = \big|\mathbb{E}\left[h(x)\right] - \mathbb{E}\left[h(y)\right]\big|.$$

Many probability metrics can be expressed as above by choosing a suitable function class $\mathcal{H}$. Examples include *Total Variation* (TV), *Kolmogorov* and *Wasserstein* metrics [GS02, CGS10]. Based on the second and the fourth derivatives of the cumulant generating

function, we define the following function classes:

$$\mathcal{H}_1 = \left\{h(x) = \Psi^{(2)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}\right\}, \qquad \mathcal{H}_2 = \left\{h(x) = \Psi^{(4)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}\right\},$$
$$\mathcal{H}_3 = \left\{h(x) = \langle v, x\rangle^2 \Psi^{(2)}(\langle x, \beta\rangle) : \beta \in \mathcal{C}, \|v\|_2 = 1\right\},$$

where $\mathcal{C} \in \mathbb{R}^p$ is a closed, convex set that is bounded by the radius R. Exact calculation of such probability metrics are often difficult. The general approach is to upper bound the distance by a more intuitive metric. In our case, we observe that $d_{\mathcal{H}_j}(x, y)$ for $j = 1, 2, 3$, can be easily upper bounded by $d_{\mathrm{TV}}(x, y)$ up to a scaling constant, when the covariates have bounded support.

In our theoretical results, we rely on projected updates onto a closed convex set $\mathcal{C}$, which are of the form

$$\hat{\beta}^{t+1} = \mathcal{P}^t_{\mathcal{C}}\left(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla_\beta l(\hat{\beta}^t)\right)$$

where the projection is defined as $\mathcal{P}^t_{\mathcal{C}}(\beta) = \mathrm{argmin}_{w\in\mathcal{C}} \frac{1}{2}\|w - \beta\|^2_{\mathbf{Q}^{t-1}}$, with $\mathcal{C}$ bounded by R. This is a special case of proximal Newton-type algorithms and further generalization is straightforward (See [LSS14]). We will further assume that the covariance matrix has full rank and its smallest eigenvalue is lower bounded by a positive constant.

3.4.2 Bounded Covariates

We have the following per-step bound for the iterates generated by the Newton-Stein method, when the covariates are supported on a ball.

Theorem 3.4.1 (Local convergence)**.** *Assume that the covariates $x_1, x_2, ..., x_n$ are i.i.d. random vectors supported on a ball of radius $\sqrt{K}$ with*

$$\mathbb{E}[x_i] = 0 \qquad \textit{and} \qquad \mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}.$$

Further assume that the cumulant generating function ϕ has bounded 2nd-5th derivatives and that the set $\mathcal{C}$ is bounded by R. For $\left\{\hat{\beta}^t\right\}_{t>0}$ given by the Newton-Stein method for $\gamma = 1$, define the event

$$\mathcal{E} = \left\{\inf_{\|u\|_2=1} \left|\mu_2(\hat{\beta}^t)\langle u, \mathbf{\Sigma} u\rangle + \mu_4(\hat{\beta}^t)\langle u, \mathbf{\Sigma}\hat{\beta}^t\rangle^2\right| > 2\kappa^{-1} \quad \forall t, \quad \beta_* \in \mathcal{C}\right\} \tag{3.7}$$

for some positive constant κ, and the optimal value β_*. If $n, |S|$ and p are sufficiently large,

then there exist constants c, c_1, c_2 depending on the radii K, R, $\mathbb{P}(\mathcal{E})$ and the bounds on $\Psi^{(2)}$ and $|\Psi^{(4)}|$ such that conditioned on the event $\mathcal{E}$, with probability at least $1 - c/p^2$, we have

$$\left\|\hat{\beta}^{t+1} - \beta_*\right\|_2 \leq \tau_1 \left\|\hat{\beta}^t - \beta_*\right\|_2 + \tau_2 \left\|\hat{\beta}^t - \beta_*\right\|_2^2, \tag{3.8}$$

where the coefficients τ_1 and τ_2 are deterministic constants defined as

$$\tau_1 = \kappa \mathbf{D}(x, z) + c_1 \kappa \sqrt{\frac{p}{\min\{p/\log(p)|S|, n/\log(n)\}}}, \qquad \tau_2 = c_2 \kappa, \tag{3.9}$$

and $\mathbf{D}(x, z)$ is defined as

$$\mathbf{D}(x, z) = \|\mathbf{\Sigma}\|_2 \; d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2 R^2 \; d_{\mathcal{H}_2}(x, z) + d_{\mathcal{H}_3}(x, z), \tag{3.10}$$

for a multivariate Gaussian random variable z with the same mean and covariance as x_i's.

The bound in Equation (3.8) holds with high probability, and the coefficients τ_1 and τ_2 are deterministic constants which will describe the convergence behavior of the Newton-Stein method. Observe that the coefficient τ_1 is sum of two terms: $\mathbf{D}(x, z)$ measures how accurate the Hessian estimation is, and the second term depends on the subsampling size $|S|$ and the data dimensions n, p.

Theorem 3.4.1 shows that the convergence of Newton-Stein method can be upper bounded by a compositely converging sequence, that is, the squared term will dominate at first providing us with a quadratic rate, then the convergence will transition into a linear phase as the iterate gets close to the optimal value. The coefficients τ_1 and τ_2 govern the linear and quadratic terms, respectively. The effect of subsampling appears in the coefficient of linear term. In theory, there is a threshold for the subsampling size $|S|$, namely $\mathcal{O}(n/\log(n))$, beyond which further subsampling has no effect. The transition point between the quadratic and the linear phases is determined by the subsampling size and the properties of the data. The phase transition behavior can be observed through the right plot in Figure 3.1.

Using the above theorem, we state the following corollary.

Corollary 3.4.2. *Assume that the assumptions of Theorem 3.4.1 hold. For a constant $\delta \geq \mathbb{P}\left(\mathcal{E}^C\right)$, and a tolerance ϵ satisfying*

$$\epsilon \geq 20R\left\{c/p^2 + \delta\right\},$$

and for an iterate satisfying $\mathbb{E}\big[\|\hat{\beta}^t - \beta_*\|_2\big] > \epsilon$*, the following inequality holds for the iterates of Newton-Stein method,*

$$\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2\right] \leq \tilde{\tau}_1 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2\right] + \tau_2 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^2\right],$$

where $\tilde{\tau}_1 = \tau_1 + 0.1$ *and* τ_1, τ_2 *are as in Theorem 3.4.1.*

The bound stated in the above corollary is an analogue of composite convergence (given in Equation (3.8)) in expectation. Note that our results make strong assumptions on the derivatives of the cumulant generating function ϕ. We emphasize that these assumptions are valid for linear and logistic regressions. An example that does not fit in our scheme is *Poisson regression* with $\phi(z) = e^z$. However, we observed empirically that the algorithm still provides significant improvement.

The following theorem characterizes the local convergence behavior of a compositely converging sequence.

Theorem 3.4.3. *Assume that the assumptions of Theorem 3.4.1 hold with* $\tau_1 < 1$ *and for* $\vartheta = \big\|\hat{\beta}^0 - \beta_*\big\|_2$ *define the interval* $\Xi = \left(\frac{\tau_1 \vartheta}{1-\tau_2 \vartheta}, \vartheta\right)$*. Conditioned on the event* $\mathcal{E} \cap \big\{\vartheta < (1-\tau_1)/\tau_2\big\}$*, there exists a constant* c *such that with probability at least* $1 - c/p^2$*, the number of iterations to reach a tolerance of* ϵ *cannot exceed*

$$\inf_{\xi \in \Xi} \mathcal{J}(\xi) \coloneqq \log_2 \left(\frac{\log\left(\tau_1 + \tau_2 \xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)} \right) + \frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2 \xi)}, \tag{3.11}$$

where the constants τ_1 *and* τ_2 *are as in Theorem 3.4.1.*

The expression in Equation 3.11 has two terms: the first one is due to the quadratic phase whereas the second one is due to the linear phase. To obtain the properties of local convergence, a locality constraint is required. We note that $\tau_1 < 1$ is a necessary assumption, which is satisfied for sufficiently large n and $|S|$.

In the following, we establish the global convergence of the Newton-Stein method coupled with a backtracking line search—which is explicitly given in Section 3.4.4.

Theorem 3.4.4 (Global Convergence)**.** *Assume that the assumptions of Theorem 3.4.1 hold and at each step, the step size* γ_t *of the Newton-Stein method is determined by the backtracking line search with parameters* a *and* b*. Then conditioned on the event* $\mathcal{E}$*, there exists a constant* c *such that with probability at least* $1 - c/p^2$*, the sequence of iterates* $\{\hat{\beta}^t\}_{t>0}$ *generated by the Newton-Stein method converges globally.*

3.4.3 Sub-Gaussian Covariates

In this section, we carry our analysis to the more general case, where the covariates are sub-Gaussian vectors.

Theorem 3.4.5 (Local convergence)**.** *Assume that $x_1, x_2, ..., x_n$ are i.i.d. sub-Gaussian random vectors with sub-Gaussian norm K such that*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}[\|x_i\|_2] = \mu \qquad \text{and} \qquad \mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}.$$

Further assume that the cumulant generating function ϕ is uniformly bounded and has bounded 2nd-5th derivatives and that $\mathcal{C}$ is bounded by R. For $\left\{\hat{\beta}^t\right\}_{t>0}$ given by the Newton-Stein method and the event $\mathcal{E}$ in Equation (3.7), if we have $n, |S|$ and p sufficiently large and

$$n^{0.2}/\log(n) \gtrsim p,$$

then there exist constants c_1, c_2, c_3, c_4 depending on the eigenvalues of $\mathbf{\Sigma}$, the radius R, μ, $\mathbb{P}(\mathcal{E})$ and the bounds on $\Psi^{(2)}$ and $|\Psi^{(4)}|$ such that conditioned on the event $\mathcal{E}$, with probability at least $1 - c_1 e^{-c_2 p}$, the bound given in Equation 3.8 holds for constants

$$\tau_1 = \kappa \mathbf{D}(x, z) + c_3 \kappa \sqrt{\frac{p}{\min\left\{|S|, n^{0.2}/\log(n)\right\}}}, \qquad \tau_2 = c_4 \kappa p^{1.5}, \tag{3.12}$$

where $\mathbf{D}(x, z)$ defined as in Equation (3.10).

The above theorem is more restrictive than Theorem 3.4.1. We require n to be much larger than the dimension p. Also note that a factor of $p^{1.5}$ appears in the coefficient of the quadratic term. We also notice that the threshold for the subsample size reduces to $n^{0.2}/\log(n)$.

We have the following analogue of Corrolary 3.4.2.

Corollary 3.4.6. *Assume that the assumptions of Theorem 3.4.5 hold. For a constant $\delta \geq \mathbb{P}\left(\mathcal{E}^C\right)$, and a tolerance ϵ satisfying*

$$\epsilon \geq 20R\sqrt{c_1 e^{-c_2 p} + \delta},$$

and for an iterate satisfying $\mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2\right] > \epsilon$, the iterates of Newton-Stein method will

satisfy,

$$\mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2\right] \leq \tilde{\tau}_1 \mathbb{E}\left[\|\hat{\beta}^{t}-\beta_*\|_2\right] + \tau_2 \mathbb{E}\left[\|\hat{\beta}^{t}-\beta_*\|_2^2\right],$$

where $\tilde{\tau}_1 = \tau_1 + 0.1$ *and* τ_1, τ_2 *are as in Theorem 3.4.5.*

When the covariates are in fact multivariate normal, we have $\mathbf{D}(x,z) = 0$ which implies that the coefficient τ_1 is smaller. Correspondingly, the quadratic phase lasts longer providing better performance.

We conclude this section by noting that the global convergence properties of the sub-Gaussian case is very similar to the previous section where we had bounded covariates.

3.4.4 Algorithm Parameters

Newton-Stein method takes two input parameters and for those, we suggest near-optimal choices based on our theoretical results. We further discuss the choice of a covariance estimation method which provides additional improvements to the proposed algorithm.

- *Subsample size:* Newton-Stein method uses a subset of indices to approximate the covariance matrix $\mathbf{\Sigma}$. Corollary 5.50 of [Ver10] proves that a sample size of $\mathcal{O}(p)$ is sufficient for sub-Gaussian covariates and that of $\mathcal{O}(p\log(p))$ is sufficient for arbitrary distributions supported in some ball to estimate a covariance matrix by its sample mean estimator. In the regime we consider, $n \gg p$, we suggest to use a sample size of $|S| = \mathcal{O}(p\log(p))$ for this task.

- *Covariance estimation method:* Many methods have been suggested to improve the estimation of the covariance matrix and almost all of them rely on the concept of *shrinkage* [CCS10, DGJ13]. Therefore, we suggest to use a thresholding based approach suggested by [EM15]. For a given threshold r, we take the largest r eigenvalues of the subsampled covariance estimator, setting rest of them to $(r+1)$-th eigenvalue. Eigenvalue thresholding can be considered as a shrinkage operation which will retain only the important second order information. Choosing the rank threshold r can be simply done on the sample mean estimator of $\mathbf{\Sigma}$. After obtaining the subsampled estimate of the mean, one can either plot the spectrum and choose manually or use an optimal technique from [DGJ13]. The suggested method requires a single time $\mathcal{O}(rp^2)$ computation and reduces the cost of inversion from $Ons(p^3)$ to $Ons(rp^2)$. We

highlight that the Newton-Stein method was originally presented with the eigenvalue thresholding in an early version of this chapter [Erd15].

- *Step size:* Step size choices for the Newton-Stein method are quite similar to those of Newton-type methods (i.e., see [BV04]). In the *damped phase*, one should use a line search algorithm such as *backtracking* with parameters $a \in (0, 0.5)$ and $b \in (0, 1)$. Defining the modified gradient (or composite gradient [LSS14]) $D_\gamma(\hat{\beta}^t) = \frac{1}{\gamma}\big\{\hat{\beta}^t - \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla l(\hat{\beta}^t))\big\}$, we compute the step size via

$$\gamma = \bar{\gamma}; \quad \textbf{while: } l\left(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)\right) > l(\hat{\beta}^t) - a\gamma\langle \nabla l(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle, \quad \gamma \leftarrow \gamma b.$$

  The above line search algorithm leads to global convergence with high probability as stated in Theorem 3.4.4.

  The step size choice for the local phase depends on the use of eigenvalue thresholding. If no shrinkage method is applied, line search algorithm should be initialized with $\bar{\gamma} = 1$. If a shrinkage method (e.g. eigenvalue thresholding) is applied, then choosing a larger local step size may provide faster convergence. If the data follows the r-spiked model, the optimal step size will be close to 1 if there is no subsampling. However, due to fluctuations resulting from subsampling, starting with $\bar{\gamma} = 1.2$ will provide faster local rates. This case has been explicitly studied in a preliminary version of this work [Erd15]. A heuristic derivation and a detailed discussion can also be found in Section B.3 in the Appendix.

3.5 Experiments

In this section, we validate the performance of Newton-Stein method through extensive numerical studies. We experimented on two commonly used GLM optimization problems, namely, *Logistic Regression* (LR) and *Linear Regression* (OLS). LR minimizes Equation (1.8) for the logistic function $\phi(z) = \log(1 + e^z)$, whereas OLS minimizes the same equation for $\phi(z) = z^2/2$. In the following, we briefly describe the algorithms that are used in the experiments:

- *Newton Method* (NM) uses the inverse Hessian evaluated at the current iterate, and may achieve local quadratic convergence. NM steps require $\mathcal{O}(np^2 + p^3)$ computation which makes it impractical for large-scale data sets.

- *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) forms a curvature matrix by cultivating the information from the iterates and the gradients at each iteration. Under certain assumptions, the convergence rate is locally super-linear and the per-iteration cost is comparable to that of first order methods.

- *Limited Memory BFGS* (L-BFGS) is similar to BFGS, and uses only the recent few iterates to construct the curvature matrix, gaining significant performance in terms of memory usage.

- *Gradient Descent* (GD) update is proportional to the negative of the full gradient evaluated at the current iterate. Under smoothness assumptions, GD achieves a locally linear convergence rate, with $\mathcal{O}(np)$ per-iteration cost.

- *Accelerated Gradient Descent* (AGD) is proposed by Nesterov [Nes83], which improves over the gradient descent by using a momentum term. Performance of AGD strongly depends of the smoothness of the function.

For all the algorithms, we use a constant step size that provides the fastest convergence. We use the Newton-Stein method with eigenvalue thresholding as described in Section 3.4.4. The parameters such as subsample size $|S|$, and rank r are selected by following the guidelines described in Section 3.4.4. The rank threshold r (which is an input to the eigenvalue thresholding) is specified at the title of each plot.

3.5.1 Simulations With Synthetic Data Sets

Synthetic data sets, S3, S10, and S20 are generated through a multivariate Gaussian distribution where the covariance matrix follows r-spiked model, i.e., $r = 3$ for S3 and $r = 20$ for S20. To generate the covariance matrix, we first generate a random orthogonal matrix, say $\mathbf{M}$. Next, we generate a diagonal matrix $\mathbf{\Lambda}$ that contains the eigenvalues, i.e., the first r diagonal entries are chosen to be large, and rest of them are equal to 1. Then, we let $\mathbf{\Sigma} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^T$. For dimensions of the data sets, see Table 3.2. We also emphasize that the data dimensions are chosen so that Newton method still does well.

The simulation results are summarized in Figure 3.2. Further details regarding the experiments can be found in Table 3.1. We observe that Newton-Stein method (NewSt) provides a significant improvement over the classical techniques.

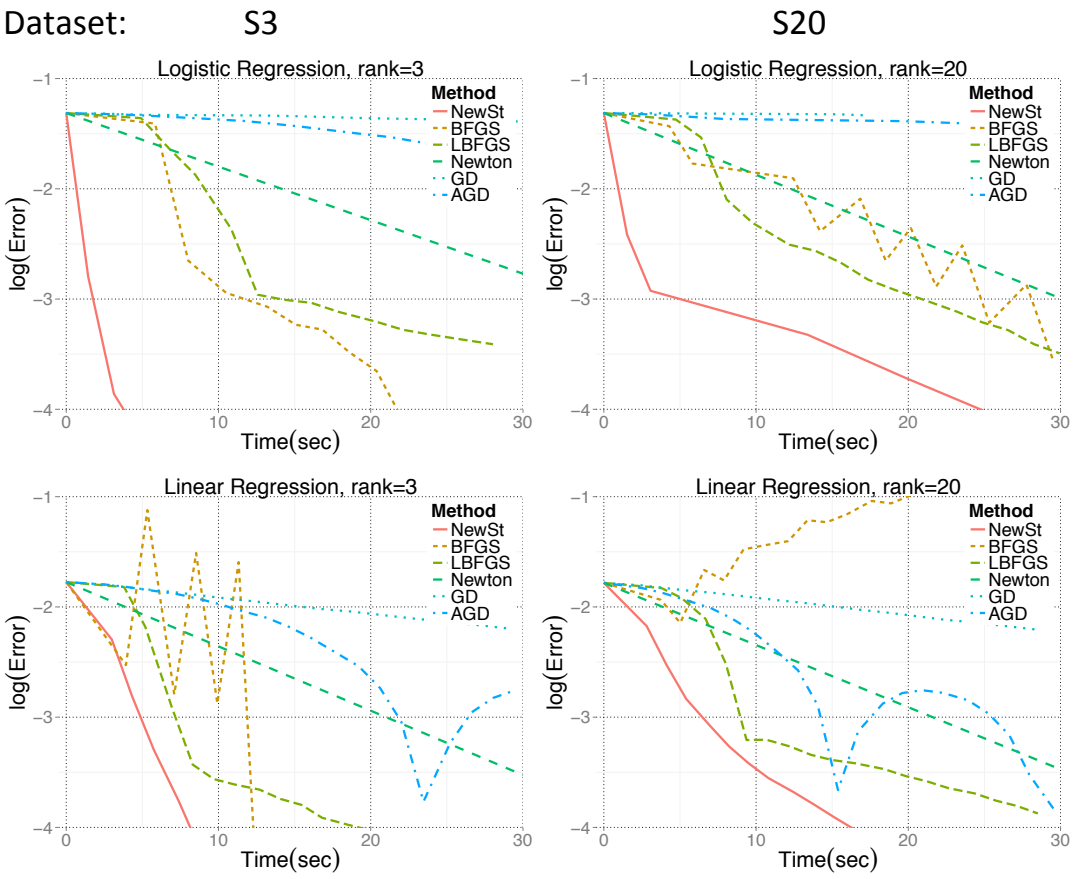

Figure 3.2: Performance of various optimization methods on two different simulated data sets. Red straight line represents the Newton-Stein method (NewSt). y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and time elapsed in seconds, respectively.

Observe that the convergence rate of NewSt has a clear phase transition point in the top left plot in Figure 3.2. As argued earlier, this point depends on various factors including subsampling size $|S|$ and data dimensions n, p, the rank threshold r and structure of the covariance matrix. The prediction of the phase transition point is an interesting line of research. However, our convergence guarantees are conservative and we believe that they cannot be used for this purpose.

3.5.2 Experiments With Real Data Sets

We experimented on two real data sets where the data sets are downloaded from UCI repository [Lic13]. Both data sets satisfy $n \gg p$, but we highlight the difference between

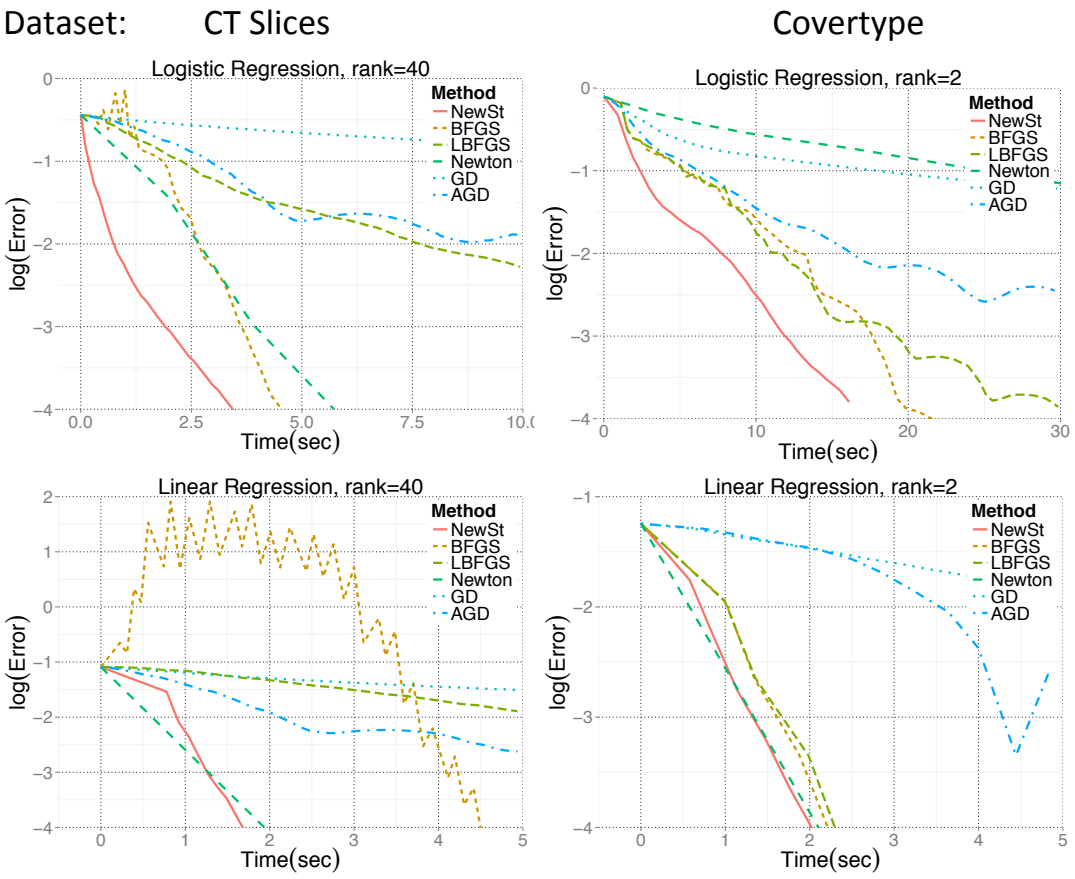

Figure 3.3: Performance of various optimization methods on two different real data sets obtained from [Lic13]. Red straight line represents the Newton-Stein method (NewSt). y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and time elapsed in seconds, respectively.

the proportions of dimensions n/p. See Table 3.2 for details.

We observe that Newton-Stein method performs better than classical methods on real data sets as well. More specifically, the methods that come closer to NewSt is Newton method for moderate n and p and BFGS when n is large.

The optimal step-size for Newton-Stein method will typically be larger than 1 which is mainly due to eigenvalue thresholding operation. This feature is desirable if one is able to obtain a large step-size that provides convergence. In such cases, the convergence is likely to be faster, yet more unstable compared to the smaller step size choices. We observed that similar to other second order algorithms, Newton-Stein method is also susceptible to the step size selection. If the data is not well-conditioned, and the subsample size is not

sufficiently large, algorithm might have poor performance. This is mainly because the subsampling operation is performed only once at the beginning. Therefore, it might be good in practice to subsample once in every few iterations.

Data set	S3				S20			
Type	LR		LS		LR		LS	
Method	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter
NewSt	10.637	2	8.763	4	23.158	4	16.475	10
BFGS	22.885	8	13.149	6	40.258	17	54.294	37
LBFGS	46.763	19	19.952	11	51.888	26	33.107	20
Newton	55.328	2	38.150	1	47.955	2	39.328	1
GD	865.119	493	155.155	100	1204.01	245	145.987	100
AGD	169.473	82	65.396	42	182.031	83	56.257	38

Data set	CT Slices				Covertype			
Type	LR		LS		LR		LS	
Method	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter	Time(sec)	Iter
NewSt	4.191	32	1.799	11	16.113	31	2.080	5
BFGS	4.638	35	4.525	37	21.916	48	2.238	3
LBFGS	26.838	217	22.679	180	30.765	69	2.321	3
Newton	5.730	3	1.937	1	122.158	40	2.164	1
GD	96.142	1156	61.526	721	194.473	446	22.738	60
AGD	96.142	880	45.864	518	80.874	186	32.563	77

Table 3.1: Details of the experiments presented in Figures 3.2 and 3.3.

3.5.3 Analysis of Number of Iterations

We provide additional plots to better understand the convergence behavior of the algorithms. Plots in Figure 3.4 show the decrease in $\log_{10}(\|\hat{\beta}^t - \beta_0\|_2)$ error over iterations (instead of time elapsed).

We observe from the plots that Newton method enjoys the fastest convergence rate as expected. The one that is closest to Newton method is the Newton-Stein method. This is because the Hessian estimator used by Newton-Stein method better approximates the true Hessian as opposed to Quasi-Newton methods. We emphasize that x axes in Figure 3.4 denote the number of iterations whereas in figures shown previously in this section x axes were the time elapsed.

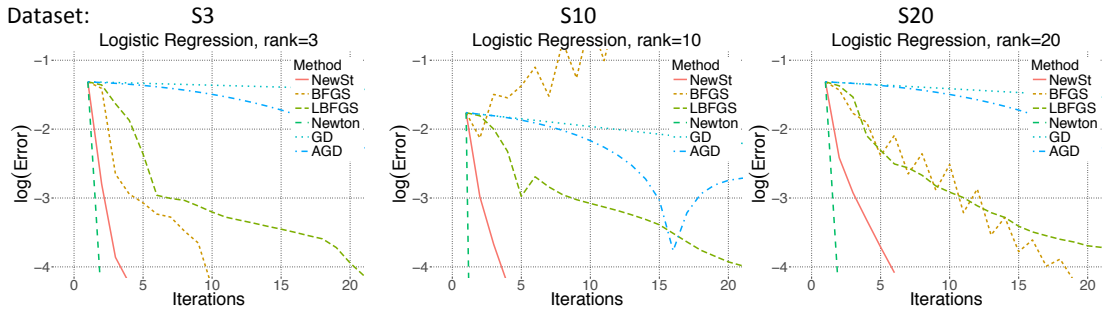

Figure 3.4: Figure shows the convergence behavior over the number of iterations. y and x axes denote $\log_{10}(\|\hat{\beta}^t - \beta_*\|_2)$ and the number iterations, respectively.

3.6 Proof of Main Results

3.6.1 Proofs of Theorems 3.4.1 and 3.4.5

We will provide the proofs of Theorems 3.4.1 and 3.4.5 in parallel as they follow from similar steps. The only difference is the application of the lemmas that are provided in the previous sections. On the event $\mathcal{E}$, we write,

$$\begin{aligned}\hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \nabla_\beta l(\hat{\beta}^t) &= \hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi (\hat{\beta}^t - \beta_*), \qquad (3.13)\\ &= \left(I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right) (\hat{\beta}^t - \beta_*)\,.\end{aligned}$$

In the following, we will work on the event that $\widehat{\boldsymbol{\Sigma}}_S$ is invertible and that $[\mathbf{Q}^t]^{-1}$ is positive definite. We later show that conditioned on $\mathcal{E}$, this event holds with very high probability when $|S|$ is sufficiently large.

Data set	n	p	Reference, UCI repo [Lic13]
CT slices	53500	386	[GKS+11]
Covertype	581012	54	[BD99]
HIGGS	11000000	28	[BSW14]
S3	500000	300	3-spiked model, [DGJ13]
S10	500000	300	10-spiked model, [DGJ13]
S20	500000	300	20-spiked model, [DGJ13]

Table 3.2: Data sets used in the experiments.

We use the nonexpensiveness of the projection $\mathcal{P}_{\mathcal{C}}^t$, i.e., for any $u, u' \in \mathbb{R}^p$ and $v = \mathcal{P}_{\mathcal{C}}^t(u)$, $v' = \mathcal{P}_{\mathcal{C}}^t(u')$ we have $\langle u - u', [\mathbf{Q}^t]^{-1}(u - u')\rangle \geq \langle v - v', [\mathbf{Q}^t]^{-1}(v - v')\rangle$. This simply means that the projection decreases the distance. Therefore, we can write

$$\begin{aligned}\left\|\hat{\beta}^{t+1} - \beta_*\right\|_{\mathbf{Q}^{t^{-1}}} &\leq \left\|\hat{\beta}^t - \beta_* - \gamma \mathbf{Q}^t \nabla_\beta l(\hat{\beta}^t)\right\|_{\mathbf{Q}^{t^{-1}}} \\ &\leq \left\|[\mathbf{Q}^t]^{-1/2} - \gamma [\mathbf{Q}^t]^{1/2} \int_0^1 \boldsymbol{\nabla}_\beta^2 l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi\right\|_2 \left\|\hat{\beta}^t - \beta_*\right\|_2 . \quad (3.14)\end{aligned}$$

The coefficient of $\|\hat{\beta}^t - \beta_*\|_2$ in Equation 3.14 determines the convergence behavior of the algorithm. Switching back to l_2 norm, we obtain an upper bounded of the form

$$\left\|\hat{\beta}^{t+1} - \beta_*\right\|_2 \leq \left\|\mathbf{Q}^t\right\|_2 \left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}_\beta^2 l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi\right\|_2 \left\|\hat{\beta}^t - \beta_*\right\|_2 ,$$

where we have set step size $\gamma = 1$. First, we will bound the second term on the right hand side. We define the following,

$$\mathfrak{E}(\beta) = \mathbb{E}\left[\Psi^{(2)}(\langle x, \beta\rangle)\right] \boldsymbol{\Sigma} + \mathbb{E}\left[\Psi^{(4)}(\langle x, \beta\rangle)\right] \boldsymbol{\Sigma} \beta \beta^T \boldsymbol{\Sigma} .$$

Note that for a function f and fixed β, $\mathbb{E}[f(\langle x, \beta\rangle)] = h(\beta)$ is a function of β. With a slight abuse of notation, we write $\mathbb{E}[f(\langle x, \hat{\beta}\rangle)] = h(\hat{\beta})$ as a random variable. We have

$$\begin{aligned}\left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}_\beta^2 l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi\right\|_2 &\leq \left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \quad (3.15) \\ &+ \left\|[\mathbb{E}[xx^T \Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \\ &+ \left\|\int_0^1 \boldsymbol{\nabla}_\beta^2 l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi - \mathbb{E}\left[xx^T \int_0^1 \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi\right]\right\|_2 \\ &+ \left\|\mathbb{E}[xx^T \Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathbb{E}\left[xx^T \int_0^1 \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi\right]\right\|_2 .\end{aligned}$$

For the first term on the right hand side, we state the following lemma.

Lemma 3.6.1. *When the covariates are sub-Gaussian, there exist constants C_1, C_2 such that, with probability at least $1 - C_1/p^2$,*

$$\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq C_2\sqrt{\frac{p}{\min\{|S|p/\log(p), n/\log(n)\}}}.$$

Similarly, when the covariates are sampled from a distribution with bounded support, there exist constants C_1', C_2', C_3' such that, with probability $1 - C_1'e^{-C_2'p}$,

$$\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq C_3'\sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}},$$

where the constants depend on K, B and the radius R.

Proof. In the following, we will only provide the proof for the bounded support case. The proof for the sub-Gaussian covariates follows from the same steps, by only replacing Lemma B.1.2 with Lemma B.1.1, and Lemma B.2.1 with Lemma B.2.3.

Using a uniform bound on the feasible set, we write

$$\begin{aligned}&\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \\ &\leq \sup_{\beta\in\mathcal{C}} \left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S + \hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mathbb{E}[\Psi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma} - \mathbb{E}[\Psi^{(4)}(\langle x,\beta\rangle)]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2.\end{aligned}$$

We will find an upper bound for the quantity inside the supremum. By denoting the expectations of $\hat{\mu}_2(\beta)$ and $\hat{\mu}_4(\beta)$, with $\mu_2(\beta)$ and $\mu_4(\beta)$ respectively, we write

$$\begin{aligned}&\left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S + \hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mathbb{E}[\Psi^{(2)}(\langle x,\beta\rangle)]\mathbf{\Sigma} - \mathbb{E}[\Psi^{(4)}(\langle x,\beta\rangle)]\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2 \\ &\leq \left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 + \left\|\hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mu_4(\beta)\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2.\end{aligned}$$

For the first term on the right hand side, we have

$$\begin{aligned}\left\|\hat{\mu}_2(\beta)\widehat{\mathbf{\Sigma}}_S - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 \leq& |\hat{\mu}_2(\beta)| \left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + \|\mathbf{\Sigma}\|_2 |\hat{\mu}_2(\beta) - \mu_2(\beta)|, \\ \leq& B_2 \left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + K |\hat{\mu}_2(\beta) - \mu_2(\beta)|.\end{aligned}$$

By the Lemmas B.1.2 and B.2.1, for an absolute constant c, we have with probability $1 - 1/p^2$,

$$\begin{aligned}\sup_{\beta\in\mathcal{C}} \left\|\hat{\mu}_2(\beta)\zeta_r(\widehat{\mathbf{\Sigma}}_S) - \mu_2(\beta)\mathbf{\Sigma}\right\|_2 \leq& B_2 c\sqrt{K\|\mathbf{\Sigma}\|_2}\sqrt{\frac{\log(p)}{|S|}} + 3B_2K\sqrt{\frac{p\log(n)}{n}},\\ \leq& 3cB_2K\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}},\\ =& \mathcal{O}\left(\sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}}\right).\end{aligned}$$

For the second term, we have

$$\begin{aligned}&\left\|\hat{\mu}_4(\beta)\widehat{\mathbf{\Sigma}}_S\beta(\widehat{\mathbf{\Sigma}}_S\beta)^T - \mu_4(\beta)\mathbf{\Sigma}\beta(\mathbf{\Sigma}\beta)^T\right\|_2\\ &\leq |\hat{\mu}_4(\beta)|\left\|\widehat{\mathbf{\Sigma}}_S\beta\beta^T\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2 + |\hat{\mu}_4(\beta) - \mu_4(\beta)|\left\|\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2,\\ &\leq B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2 + \|\mathbf{\Sigma}\|_2\right\}\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + R^2\|\mathbf{\Sigma}\|_2^2|\hat{\mu}_4(\beta) - \mu_4(\beta)|,\\ &\leq B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2 + K\right\}\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 + R^2K^2|\hat{\mu}_4(\beta) - \mu_4(\beta)|.\end{aligned}$$

Again, by the Lemmas B.1.2 and B.2.1, for an absolute constant c, we have with probability $1 - 1/p^2$,

$$\begin{aligned}B_4R^2\left\{\|\widehat{\mathbf{\Sigma}}_S\|_2 + K\right\}\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq& cKB_4R^2\left\{2K + cK\sqrt{\frac{\log(p)}{|S|}}\right\}\sqrt{\frac{\log(p)}{|S|}},\\ \leq& 2cK^2B_4R^2\sqrt{\frac{\log(p)}{|S|}} + c^2K^2B_4R^2\frac{\log(p)}{|S|},\\ \leq& 2cK^2B_4R^2\left(1 + c\sqrt{\frac{\log(p)}{|S|}}\right)\sqrt{\frac{\log(p)}{|S|}},\\ \leq& 4cK^2B_4R^2\sqrt{\frac{\log(p)}{|S|}},\\ =& \mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right),\end{aligned}$$

for sufficiently large $|S|$, i.e., $|S| \geq c^2 \log(p)$.

Further, by Lemma B.2.1, we have with probability $1 - 2e^{-p}$,

$$\sup_{\beta \in \mathcal{C}} |\hat{\mu}_4(\beta) - \mu_4(\beta)| \leq 3B_4 \sqrt{\frac{p \log(n)}{n}} = \mathcal{O}\left(\sqrt{\frac{p \log(n)}{n}}\right).$$

Combining the above results, for sufficiently large $p, |S|$, we have with probability at least $1 - 1/p^2 - 2e^{-p}$,

$$\begin{aligned}
&\sup_{\beta \in \mathcal{C}} \left\| \hat{\mu}_2(\beta) \zeta_r(\widehat{\mathbf{\Sigma}}_S) - \mu_2(\beta) \mathbf{\Sigma} \right\|_2 + \sup_{\beta \in \mathcal{C}} \left\| \hat{\mu}_4(\beta) \widehat{\mathbf{\Sigma}}_S \beta (\widehat{\mathbf{\Sigma}}_S \beta)^T - \mu_4(\beta) \mathbf{\Sigma} \beta (\mathbf{\Sigma} \beta)^T \right\|_2 \\
&\leq 3B_2 K c \sqrt{\frac{p}{\min\{p/\log(p)|S|, n/\log(n)\}}} \\
&\quad + 4cK^2 B_4 R^2 \sqrt{\frac{\log(p)}{|S|}} + 3B_4 R^2 K^2 \sqrt{\frac{p \log(n)}{n}}, \\
&\leq 3B_2 K c \sqrt{\frac{p}{\min\{p/\log(p)|S|, n/\log(n)\}}} \\
&\quad + 4cK^2 B_4 R^2 \sqrt{\frac{p}{\min\{p/\log(p)|S|, n/\log(n)\}}}, \\
&\leq CK \max\{B_2, B_4 K R^2\} \sqrt{\frac{p}{\min\{p/\log(p)|S|, n/\log(n)\}}}, \\
&= \mathcal{O}\left(\sqrt{\frac{p}{\min\{|S|p/\log(p), n/\log(n)\}}}\right).
\end{aligned}$$

Hence, for some constants C_1, C_2, with probability $1 - C_1/p^2$, we have

$$\left\| [\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\beta}^t) \right\|_2 \leq C_2 \sqrt{\frac{p}{\min\{|S|p/\log(p), n/\log(n)\}}},$$

where the constants depend on $K, B = \max\{B_2, B_4\}$ and the radius R. □

Lemma 3.6.2. *The bias term can be upper bounded by*

$$\left\| \mathbb{E}[x x^T \Psi^{(2)}(\langle x, \hat{\beta}^t \rangle)] - \mathfrak{E}(\hat{\beta}^t) \right\|_2 \leq d_{\mathcal{H}_3}(x, z) + \|\mathbf{\Sigma}\|_2 \; d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2 R^2 \; d_{\mathcal{H}_2}(x, z),$$

for both sub-Gaussian and bounded support cases.

Proof. For a random variable $z \sim \mathsf{N}_p(0, \mathbf{\Sigma})$, by the triangle inequality, we write

$$\begin{aligned}
&\left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \\
&\leq \left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathbb{E}[zz^T\Psi^{(2)}(\langle z, \hat{\beta}^t\rangle)]\right\|_2 + \left\|\mathbb{E}[zz^T\Psi^{(2)}(\langle z, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2
\end{aligned}$$

For the first term on the right hand side, we have

$$\begin{aligned}
&\left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathbb{E}[zz^T\Psi^{(2)}(\langle z, \hat{\beta}^t\rangle)]\right\|_2 \\
&\leq \sup_{\beta\in\mathcal{C}} \sup_{\|v\|_2=1} \left|\mathbb{E}\left[\langle v, x\rangle^2\Psi^{(2)}(\langle x, \beta\rangle)\right] - \mathbb{E}\left[\langle v, z\rangle^2\Psi^{(2)}(\langle z, \beta\rangle)\right]\right|, \\
&\leq d_{\mathcal{H}_3}(x, z).
\end{aligned}$$

For the second term, we write

$$\begin{aligned}
&\left\|[\mathbb{E}[zz^T\Psi^{(2)}(\langle z, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \\
&\leq \sup_{\beta\in\mathcal{C}} \left\|\mathbb{E}[zz^T\Psi^{(2)}(\langle z, \beta\rangle)] - \mathbb{E}[\Psi^{(2)}(\langle x, \beta\rangle)]\mathbf{\Sigma} + \mathbb{E}\left[\Psi^{(4)}(\langle x, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2, \\
&\leq \sup_{\beta\in\mathcal{C}} \Big\|\mathbb{E}[\Psi^{(2)}(\langle z, \beta\rangle)]\mathbf{\Sigma} + \mathbb{E}\left[\Psi^{(4)}(\langle z, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma} \\
&\qquad\qquad - \mathbb{E}[\Psi^{(2)}(\langle x, \beta\rangle)]\mathbf{\Sigma} - \mathbb{E}\left[\Psi^{(4)}(\langle x, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\Big\|_2, \\
&\leq \sup_{\beta\in\mathcal{C}} \left\|\mathbb{E}[\Psi^{(2)}(\langle z, \beta\rangle)]\mathbf{\Sigma} - \mathbb{E}[\Psi^{(2)}(\langle x, \beta\rangle)]\mathbf{\Sigma}\right\|_2, \\
&\qquad + \sup_{\beta\in\mathcal{C}} \left\|\mathbb{E}\left[\Psi^{(4)}(\langle z, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma} - \mathbb{E}\left[\Psi^{(4)}(\langle x, \beta\rangle)\right]\mathbf{\Sigma}\beta\beta^T\mathbf{\Sigma}\right\|_2, \\
&\leq \|\mathbf{\Sigma}\|_2 \sup_{\beta\in\mathcal{C}} \left|\mathbb{E}[\Psi^{(2)}(\langle z, \beta\rangle)] - \mathbb{E}[\Psi^{(2)}(\langle x, \beta\rangle)]\right| \\
&\qquad + \|\mathbf{\Sigma}\|_2^2 R^2 \sup_{\beta\in\mathcal{C}} \left|\mathbb{E}[\Psi^{(4)}(\langle z, \beta\rangle)] - \mathbb{E}[\Psi^{(4)}(\langle x, \beta\rangle)]\right|, \\
&\leq \|\mathbf{\Sigma}\|_2 d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2 R^2 d_{\mathcal{H}_2}(x, z).
\end{aligned}$$

Hence, we conclude that

$$\left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x, \hat{\beta}^t\rangle)] - \mathfrak{E}(\hat{\beta}^t)\right\|_2 \leq d_{\mathcal{H}_3}(x, z) + \|\mathbf{\Sigma}\|_2\ d_{\mathcal{H}_1}(x, z) + \|\mathbf{\Sigma}\|_2^2 R^2\ d_{\mathcal{H}_2}(x, z).$$

□

Lemma 3.6.3. *There exists constants c_1, c_2, c_3 depending on the eigenvalues of $\mathbf{\Sigma}$, B, L and R such that, with probability at least $1 - c_2 e^{-c_3 p}$*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \int_0^1 \Psi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi - \mathbb{E}\left[x x^T \int_0^1 \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi \right] \right\|_2 \leq \delta$$

where $\delta = c_1 \sqrt{\frac{p}{n^{0.2}} \log(n)}$ for sub-Gaussian covariates, and $\delta = c_1 \sqrt{\frac{p}{n} \log(n)}$ for covariates with bounded support.

Proof. We provide the proof for bounded support case. The proof for sub-Gaussian case can be carried by replacing Lemma B.2.2 with Lemma B.2.5.

By the Fubini's theorem, we have

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \int_0^1 \Psi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi - \mathbb{E}\left[x x^T \int_0^1 \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) d\xi \right] \right\|_2 \\
&= \left\| \int_0^1 \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \Psi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) - \mathbb{E}\left[x x^T \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) \right] \right\} d\xi \right\|_2 \\
&\leq \int_0^1 \left\| \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \Psi^{(2)}(\langle x_i, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) - \mathbb{E}\left[x x^T \Psi^{(2)}(\langle x, \beta_* + \xi(\hat{\beta}^t - \beta_*)\rangle) \right] \right\} \right\|_2 d\xi \\
&\leq \sup_{\beta \in \mathcal{C}} \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \Psi^{(2)}(\langle x_i, \beta\rangle) - \mathbb{E}\left[x x^T \Psi^{(2)}(\langle x, \beta\rangle) \right] \right\|_2 .
\end{aligned}$$

Using the properties of operator norm, the above bound can be written as

$$\begin{aligned}
&\sup_{\beta \in \mathcal{C}} \left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \Psi^{(2)}(\langle x_i, \beta\rangle) - \mathbb{E}\left[x x^T \Psi^{(2)}(\langle x, \beta\rangle) \right] \right\|_2 \\
&= \sup_{\beta \in \mathcal{C}} \sup_{v \in S^{p-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \Psi^{(2)}(\langle x_i, \beta\rangle) \langle x_i, v\rangle^2 - \mathbb{E}\left[\Psi^{(2)}(\langle x, \beta\rangle) \langle x, v\rangle^2 \right] \right|,
\end{aligned}$$

where S^{p-1} denotes the p-dimensional unit sphere.

For $\Delta = 0.25$, let T_Δ be an Δ-net over S^{p-1}. Using Lemma B.4.4, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\sup_{v\in S^{p-1}}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle)\langle x,v\rangle^2\right]\right|>\epsilon\right),\\
&\leq\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\sup_{v\in T_\Delta}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle)\langle x,v\rangle^2\right]\right|>\epsilon/2\right),\\
&\leq|T_\Delta|\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle)\langle x,v\rangle^2\right]\right|>\epsilon/2\right),\\
&=9^p\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}\left[\Psi^{(2)}(\langle x,\beta\rangle)\langle x,v\rangle^2\right]\right|>\epsilon/2\right).
\end{aligned}$$

By applying Lemma B.2.2 to the last line above, we obtain

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}[\Psi^{(2)}(\langle x,\beta\rangle)\langle x,v\rangle^2]\right|>4B_2K\sqrt{\frac{p}{n}\log(n)}\right)\leq 2e^{-3.2p}.$$

Notice that $3.2-\log(9)>1$. Therefore, by choosing n large enough, on the set $\mathcal{E}$, we obtain that with probability at least $1-2e^{-p}$

$$\sup_{\beta\in\mathcal{C}}\left\|\frac{1}{n}\sum_{i=1}^{n}x_ix_i^T\Psi^{(2)}(\langle x_i,\beta\rangle)-\mathbb{E}\left[xx^T\Psi^{(2)}(\langle x,\beta\rangle)\right]\right\|_2\leq 8B_2K\sqrt{\frac{p}{n}\log(n)}.$$

□

Lemma 3.6.4. *There exists a constant C depending on K and L such that,*

$$\left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x,\hat{\beta}^t\rangle)]-\mathbb{E}\left[xx^T\int_0^1\Psi^{(2)}(\langle x,\beta_*+\xi(\hat{\beta}^t-\beta_*)\rangle)d\xi\right]\right\|_2\leq\tilde{C}\|\hat{\beta}^t-\beta_*\|_2,$$

where $\tilde{C}=C$ for the bounded support case and $\tilde{C}=Cp^{1.5}$ for the sub-Gaussian case.

Proof. By the Fubini's theorem, we write

$$\left\|\mathbb{E}[xx^T\Psi^{(2)}(\langle x,\hat{\beta}^t\rangle)] - \mathbb{E}\left[xx^T\int_0^1 \Psi^{(2)}(\langle x,\beta_* + \xi(\hat{\beta}^t-\beta_*)\rangle)d\xi\right]\right\|_2,$$
$$= \left\|\int_0^1 \mathbb{E}\left[xx^T\left\{\Psi^{(2)}(\langle x,\hat{\beta}^t\rangle) - \Psi^{(2)}(\langle x,\beta_* + \xi(\hat{\beta}^t-\beta_*)\rangle)\right\}\right]d\xi\right\|_2.$$

Moving the integration out, right hand side of the above equation is smaller than

$$\int_0^1 \left\|\mathbb{E}\left[xx^T\left\{\Psi^{(2)}(\langle x,\hat{\beta}^t\rangle) - \Psi^{(2)}(\langle x,\beta_* + \xi(\hat{\beta}^t-\beta_*)\rangle)\right\}\right]\right\|_2 d\xi,$$
$$\leq \int_0^1 \left\|\mathbb{E}\left[xx^T L|\langle x,(1-\xi)(\hat{\beta}^t-\beta_*)\rangle|\right]\right\|_2 d\xi,$$
$$\leq \mathbb{E}\left[\|x\|_2^3\|\hat{\beta}^t-\beta_*\|_2\right] L\int_0^1(1-\xi)d\xi,$$
$$= \frac{L\mathbb{E}[\|x\|_2^3]}{2}\|\hat{\beta}^t-\beta_*\|_2.$$

We observe that, when the covariates are supported in the ball of radius $\sqrt{K}$, we have $\mathbb{E}[\|x\|_2^3] \leq K^{3/2}$. When they are sub-Gaussian random variables with norm K, we have $\mathbb{E}[\|x\|_2^3] \leq K^3 6^{1.5} p^{1.5}$.

□

By combining the above results, for bounded covariates we obtain

$$\left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t-\beta_*))d\xi\right\|_2$$
$$\leq \mathbf{D}(x,z) + c_1\sqrt{\frac{p}{\min\{|S|p/\log(p), n/\log(n)\}}} + c_2\|\hat{\beta}^t-\beta_*\|_2\,,$$

and for sub-Gaussian covariates, we obtain

$$\left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t-\beta_*))d\xi\right\|_2$$
$$\leq \mathbf{D}(x,z) + c_1\sqrt{\frac{p}{\min\{|S|, n^{0.2}/\log(n)\}}} + c_2 p^{1.5}\|\hat{\beta}^t-\beta_*\|_2\,,$$

where

$$\mathbf{D}(x,z) = d_{\mathcal{H}_3}(x,z) + \|\boldsymbol{\Sigma}\|_2\; d_{\mathcal{H}_1}(x,z) + \|\boldsymbol{\Sigma}\|_2^2 R^2\; d_{\mathcal{H}_2}(x,z)\,.$$

In the following, we will derive an upper bound for $\left\|\mathbf{Q}^t\right\|_2$, which is equivalent to proving the positive definiteness of $[\mathbf{Q}^t]^{-1}$ and finding a lower bound for $\|[\mathbf{Q}^t]^{-1}\|_2$. The sub-Gaussian case is more restrictive than the bounded support case. Therefore we derive the bound for the sub-Gaussian case. We have

$$\begin{aligned}\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) &= \inf_{\|u\|_2=1}\left\{\hat{\mu}_2(\hat{\beta}^t)\langle u, \widehat{\mathbf{\Sigma}}_S u\rangle + \hat{\mu}_4(\hat{\beta}^t)\langle u, \widehat{\mathbf{\Sigma}}_S\hat{\beta}^t\rangle^2\right\},\\ &\geq \inf_{\|u\|_2=1}\left\{\hat{\mu}_2(\hat{\beta}^t)\langle u, \mathbf{\Sigma} u\rangle + \hat{\mu}_4(\hat{\beta}^t)\langle u, \mathbf{\Sigma}\hat{\beta}^t\rangle^2\right\}\\ &\quad - B_2\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\|_2 - B_4R^2\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\|_2\|\widehat{\mathbf{\Sigma}}_S + \mathbf{\Sigma}\|_2.\end{aligned}$$

On the event $\mathcal{E}$, the first term on the right hand side is lower bounded by κ^{-1}. For the other terms, we use Lemma B.1.1 and write

$$\begin{aligned}\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) &\leq 2\kappa^{-1} - \|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\|_2\left\{B_2 + B_4R^2\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\|_2 + 2B_4R^2\|\mathbf{\Sigma}\|_2\right\},\\ &\leq 2\kappa^{-1} - C\sqrt{\frac{p}{|S|}}\left\{B_2 + B_4R^2C\sqrt{\frac{p}{|S|}} + 2B_4R^2\|\mathbf{\Sigma}\|_2\right\}\end{aligned}$$

with probability $1-2e^{-cp}$. When $|S| > 4pC^2\max\left\{1, 2C(B_2 + 3B_4R^2\lambda_{\max}(\mathbf{\Sigma}))\kappa\right\}^2$, with probability $1-2e^{-cp}$, we obtain

$$\lambda_{\min}\left([\mathbf{Q}^t]^{-1}\right) \geq \kappa^{-1}.$$

This proves that, with high probability, on the event $\mathcal{E}$, $[\mathbf{Q}^t]^{-1}$ is positive definite and consequently we obtain

$$\|\mathbf{Q}^t\|_2 \leq \kappa.$$

Finally, we take into account the conditioning on the event $\mathcal{E}$. Since we worked on the event $\mathcal{E}$, the probability of a desired outcome is at least $\mathbb{P}(\mathcal{E}) - \delta$, where δ is either c/p^2 or ce^{-p} depending on the distribution of the covariates. Hence, conditioned on the event $\mathcal{E}$, the probability becomes $1 - \delta/\mathbb{P}(\mathcal{E})$, which completes the proof.

3.6.2 Proofs of Corollaries 3.4.2 and 3.4.6

In the following, we provide the proof for Corollary 3.4.2. The proof for Corollary 3.4.6 follows from the exact same steps.

The statement of Theorem 3.4.1 holds on the probability space with a probability lower bounded by $\mathbb{P}(\mathcal{E}) - c/p^2$ for some constant c (See previous section). Let $\mathcal{Q}$ denote this set, on which the statement of the theorem holds without the conditioning on the event $\mathcal{E}$. Note that $\mathcal{Q} \subset \mathcal{E}$ and we also have

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{Q}) \geq \mathbb{P}(\mathcal{E}) - c/p^2. \tag{3.16}$$

This suggests that the difference between $\mathcal{Q}$ and $\mathcal{E}$ is small. By taking expectations on both sides over the set $\mathcal{Q}$, we obtain,

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\beta}^{t+1} - \beta_*\|_2; \mathcal{Q}\right] \\
&\leq \kappa \left\{ \mathbf{D}(x, z) + c_1 \sqrt{\frac{p}{\min\left\{p/\log(p)|S|, n/\log(n)\right\}}} \right\} \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2\right] \\
&+ \kappa c_2 \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^2\right]
\end{aligned}$$

where we used

$$\mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^l; \mathcal{Q}\right] \leq \mathbb{E}\left[\|\hat{\beta}^t - \beta_*\|_2^l\right], \qquad l = 1, 2.$$

Similarly for the iterate $\hat{\beta}^{t+1}$, we write

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2\right] =& \mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}\right]+\mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}^C\right],\\
\leq& \mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}\right]+2R\mathbb{P}(\mathcal{Q}^C),\\
\leq& \mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}\right]+2R\left(\mathbb{P}(\mathcal{E}^C)+\frac{c}{p^2}\right),\\
\leq& \mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}\right]+\frac{\epsilon}{10},\\
\leq& \mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2;\mathcal{Q}\right]+\frac{\mathbb{E}\left[\|\hat{\beta}^{t}-\beta_*\|_2\right]}{10}.
\end{aligned}$$

Combining these two inequalities, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\beta}^{t+1}-\beta_*\|_2\right]\\
&\leq\left\{0.1+\kappa\mathbf{D}(x,z)+c_1\kappa\sqrt{\frac{p}{\min\left\{p/\log(p)|S|,n/\log(n)\right\}}}\right\}\mathbb{E}\left[\|\hat{\beta}^{t}-\beta_*\|_2\right]\\
&+c_2\kappa\mathbb{E}\left[\|\hat{\beta}^{t}-\beta_*\|_2^2\right].
\end{aligned}$$

Hence the proof follows.

3.6.3 Proof of Theorem 3.4.3

The iterates generated by the Newton-Stein method satisfy the following inequality,

$$\|\hat{\beta}^{t+1}-\beta_*\|_2\leq\left(\tau_1+\tau_2\|\hat{\beta}^{t}-\beta_*\|_2\right)\|\hat{\beta}^{t}-\beta_*\|_2,$$

on the event $\mathcal{Q}$ where $\mathcal{Q}$ is defined in the previous section. We have observed that $\mathbb{P}(\mathcal{Q})\geq\mathbb{P}(\mathcal{E})-c/p^2$ in Equation 3.16. Since the coefficients τ_1 and τ_2 are obtained by uniform bounds on the feasible set, the above inequality holds for every t on $\mathcal{Q}$. On the event we consider, $\mathcal{Q}\cap\{\vartheta<(1-\tau_1)/\tau_2\}$, the starting point satisfies the following

$$\tau_1+\tau_2\|\hat{\beta}^{0}-\beta_*\|_2<1, \tag{3.17}$$

which implies that the sequence of iterates converges. Let $\xi \in (\epsilon, \vartheta)$ and t_ξ be the last iteration that $\|\hat{\beta}^t - \beta_*\|_2 > \xi$. Then, for $t > t_\xi$

$$\begin{aligned}\|\hat{\beta}^{t+1} - \beta_*\|_2 &\leq \left(\tau_1 + \tau_2\|\hat{\beta}^t - \beta_*\|_2\right)\|\hat{\beta}^t - \beta_*\|_2, \\ &\leq (\tau_1 + \tau_2\xi)\,\|\hat{\beta}^t - \beta_*\|_2.\end{aligned}$$

This convergence behavior describes a linear rate and requires at most

$$\frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2\xi)}$$

iterations to reach a tolerance of ϵ. For $t \leq t_\xi$, we have

$$\begin{aligned}\|\hat{\beta}^{t+1} - \beta_*\|_2 &\leq \left(\tau_1 + \tau_2\|\hat{\beta}^t - \beta_*\|_2\right)\|\hat{\beta}^t - \beta_*\|_2, \\ &\leq (\tau_1/\xi + \tau_2)\,\|\hat{\beta}^t - \beta_*\|_2^2.\end{aligned}$$

This describes a quadratic rate and the number of iterations to reach a tolerance of ξ can be upper bounded by

$$\log_2\left(\frac{\log\left(\xi\left(\tau_1/\xi + \tau_2\right)\right)}{\log\left(\tau_1/\xi + \tau_2\right)\|\hat{\beta}^0 - \beta_*\|_2}\right) \leq \log_2\left(\frac{\log\left(\tau_1 + \tau_2\xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)}\right).$$

Therefore, the overall number of iterations to reach a tolerance of ϵ is upper bounded by

$$\log_2\left(\frac{\log\left(\tau_1 + \tau_2\xi\right)}{\log\left((\tau_1/\xi + \tau_2)(1-\tau_1)/\tau_2\right)}\right) + \frac{\log(\epsilon/\xi)}{\log(\tau_1 + \tau_2\xi)}$$

which is a function of ξ. Therefore, we take the minimum over the feasible set and conclude that on $\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\}$, the number of iterations to reach a tolerance of ϵ is upper bounded by $\inf_\xi \mathcal{J}(\xi)$ with a bad event probability of c/p^2. By conditioning on the event $\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\}$, we conclude that with probability at least $1 - c'/p^2$, the statement of the theorem holds for $c' = c/\mathbb{P}(\mathcal{E} \cap \{\vartheta < (1-\tau_1)/\tau_2\})$.

3.6.4 Proof of Theorem 3.4.4

We have the following projected updates

$$\hat{\beta}^{t+1} = \mathcal{P}_{\mathcal{C}}\left(\hat{\beta}^t - \gamma_t \mathbf{Q}^t \nabla l(\hat{\beta}^t); \mathbf{Q}^t\right) = \hat{\beta}^t - \gamma_t D_{\gamma_t}(\hat{\beta}^t),$$

where we define

$$D_\gamma(\hat{\beta}^t) = \frac{1}{\gamma}\left(\hat{\beta}^t - \mathcal{P}_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla l(\hat{\beta}^t); \mathbf{Q}^t)\right).$$

For simplicity, we only consider the projection onto a convex set, i.e.,

$$\begin{aligned}\mathcal{P}^t_{\mathcal{C}}(\beta^+) = \mathcal{P}_{\mathcal{C}}(\beta^+; \mathbf{Q}^t) =& \underset{w\in\mathcal{C}}{\operatorname{argmin}} \frac{1}{2}\|w - \beta^+\|^2_{\mathbf{Q}^{t-1}}, \\ =& \underset{w\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|w - \beta^+\|^2_{\mathbf{Q}^{t-1}} + \mathbb{1}_{\mathcal{C}}(w),\end{aligned} \tag{3.18}$$

where $\mathbb{1}_{\mathcal{C}}(w)$ is the indicator function for the convex set $\mathcal{C}$, i.e.

$$\mathbb{1}_{\mathcal{C}}(w) = \begin{cases} 0 & \text{if } w \in \mathcal{C}, \\ \infty & \text{otherwise.} \end{cases}$$

We note that other projection methods (such as proximal mappings) are also applicable to our update rule.

Defining the decrement $\lambda^t = \langle \nabla l(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle$, we consider the following form of backtracking line search with update parameters $a \in (0, 0.5)$ and $b \in (0, 1)$:

$$\gamma = \bar{\gamma}; \quad \textbf{while: } l\left(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)\right) > l(\hat{\beta}^t) - a\gamma\lambda^t, \quad \gamma \leftarrow \gamma b.$$

Depending on the projection choice, there are various other search methods that can be applied. Before we move on to the convergence analysis, we first establish some properties of the modified gradient D_γ.

For a given point $w \in \mathcal{C}$, the sub-differential of the indicator function is the normal cone. This together with Equation 3.18 implies that

$$\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla l(\hat{\beta}^t) - \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla l(\hat{\beta}^t)) \in \mathbf{Q}^t\ \partial \mathbb{1}_{\mathcal{C}}(\mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma \mathbf{Q}^t \nabla l(\hat{\beta}^t))),$$

which in turn implies

$$\gamma[\mathbf{Q}^t]^{-1}\left\{D_\gamma(\hat{\beta}^t) - \mathbf{Q}^t\nabla l(\hat{\beta}^t),\right\} \in \partial \mathbb{1}_{\mathcal{C}}(\mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma\mathbf{Q}^t\nabla l(\hat{\beta}^t))),$$

and correspondingly for any $\beta \in \mathcal{C}$

$$\langle [\mathbf{Q}^t]^{-1}D_\gamma(\hat{\beta}^t) - \nabla l(\hat{\beta}^t), \mathcal{P}^t_{\mathcal{C}}(\hat{\beta}^t - \gamma\mathbf{Q}^t\nabla l(\hat{\beta}^t)) - \beta\rangle \geq 0.$$

For $\beta = \hat{\beta}^t \in \mathcal{C}$, this yields

$$\kappa^{-1}\|D_\gamma(\hat{\beta}^t)\|_2^2 \leq \langle D_\gamma(\hat{\beta}^t), [\mathbf{Q}^t]^{-1}D_\gamma(\hat{\beta}^t)\rangle \leq \langle \nabla l(\hat{\beta}^t), D_{\gamma_t}(\hat{\beta}^t)\rangle, \tag{3.19}$$

with probability at least $P(\mathcal{E}) - c/p^2$. Also note that the Hessian of the GLM problem can be upper bounded by

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^T \Psi^{(2)}(\langle x_i, \hat{\beta}^t)\right\|_2 \leq B_2 \left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right\|_2 \leq B_2 K.$$

Now we move to the convergence analysis. For a step size γ, by the convexity of the negative log-likelihood, we can write almost surely

$$\begin{aligned} l(\hat{\beta}^t - \gamma D_\gamma(\hat{\beta}^t)) \leq & l(\hat{\beta}^t) - \gamma\langle \nabla l(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle + \frac{\gamma^2 B_2 K}{2}\|D_\gamma(\hat{\beta}^t)\|_2^2, \\ \leq & l(\hat{\beta}^t) - \gamma\langle \nabla l(\hat{\beta}^t), D_\gamma(\hat{\beta}^t)\rangle \left\{1 - \frac{\gamma}{2}B_2 K\kappa\right\} \end{aligned}$$

and notice that the exit condition for the backtracking line search algorithm is satisfied when $\gamma \leq (\kappa B_2 K)^{-1}$. Hence, the line search returns a step size satisfying

$$\gamma_t \geq \min\{\bar{\gamma}, b/(\kappa B_2 K)\}.$$

Using the line search condition, we have

$$l\left(\hat{\beta}^t - \gamma_t D_{\gamma_t}(\hat{\beta}^t)\right) - l(\hat{\beta}^t) \leq -a\gamma_t\lambda^t,$$

with probability at least $\mathbb{P}(\mathcal{E}) - c/p^2$ which implies that the sequence $\{l(\hat{\beta}^t)\}_t$ is decreasing. We note that this event is independent of the iteration number due to uniform positive definite condition given in $\mathcal{E}$. Since l is continuous and $\mathcal{C}$ is closed, l is a closed function.

Hence, the sequence $\{l(\hat{\beta}^t)\}_t$ must converge to a limit. This implies that $a\gamma_t\lambda^t \to 0$. But we have $a > 0$ and $\gamma_t > \min\{\bar{\gamma}, b/(\kappa B_2 K)\} > 0$. Therefore, we conclude that $\lambda^t \to 0$. Using the inequality provided in Equation 3.19, we conclude that $\|D_\gamma(\hat{\beta}^t)\|_2$ converges to 0 which implies that the algorithm converges with probability at least $1 - \frac{c}{\mathbb{P}(\mathcal{E})}p^{-2}$, where in the last step we conditioned on $\mathcal{E}$.

3.7 Discussion

In this chapter, we proposed an efficient algorithm for training GLMs. We call our algorithm Newton-Stein method (NewSt) as it takes a Newton-type step at each iteration relying on a Stein-type lemma. The algorithm requires a one time $\mathcal{O}(|S|p^2)$ cost to estimate the covariance structure and $\mathcal{O}(np)$ per-iteration cost to form the update equations. We observe that the convergence of Newton-Stein method has a phase transition from quadratic rate to linear rate. This observation is justified theoretically along with several other guarantees for the bounded as well as the sub-Gaussian covariates such as per-step convergence bounds, conditions for local rates and global convergence with line search, etc. Parameter selection guidelines of Newton-Stein method are based on our theoretical results. Our experiments show that Newton-Stein method provides significant improvement over the classical optimization methods.

Relaxing some of the theoretical constraints is an interesting line of research. In particular, strong assumptions on the cumulant generating functions might be loosened. Another interesting direction is to determine when the phase transition point occurs, which would provide a better understanding of the effects of subsampling and eigenvalue thresholding.

Chapter 4

Subsampled Newton Methods

Contents of this chapter are based on the paper [EM15]. In this chapter, we consider the problem of minimizing a sum of n functions via projected iterations onto a convex parameter set $\mathcal{C} \subset \mathbb{R}^p$, where $n \gg p \gg 1$. In this regime, algorithms which utilize subsampling techniques are known to be effective. In this chapter, we use subsampling techniques together with low-rank approximation to design a new randomized batch algorithm which possesses comparable convergence rate to Newton method, yet has much smaller per-iteration cost. Our theoretical results can be used to obtain convergence rates of previously proposed subsampling based algorithms as well.

4.1 Introduction

In this chapter, we consider the more general problem of minimizing an average of n functions $f_i : \mathbb{R}^p \to \mathbb{R}$,

$$\underset{\beta}{\text{minimize}}\ f(\beta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(\beta)\,, \tag{4.1}$$

in a batch setting, where n is assumed to be much larger than p. Most machine learning models can be expressed as above, where each function f_i corresponds to an observation. Examples include logistic regression, support vector machines, neural networks and graphical models.

Many optimization algorithms have been developed to solve the above minimization problem using iterative methods [Bis95, BV04, Nes13]. In this chapter, we consider the

iterations of the following form

$$\beta^{t+1} = \beta^t - \gamma_t \mathbf{Q}^t \nabla_\beta f(\beta^t), \tag{4.2}$$

where γ_t is the step size and $\mathbf{Q}^t$ is a suitable scaling matrix that provides curvature information (For simplicity, we drop the projection throughout the introduction, i.e., we assume $\mathcal{C} = \mathbb{R}^p$).

Updates of the form Equation (4.2) have been extensively studied in the optimization literature. We have seen first order, second order, and Quasi-Newton methods using the iterations of the above form in Section 1.2. An alternative approach is to use *subsampling* techniques, where scaling matrix $\mathbf{Q}^t$ is based on randomly selected set of data points [Mar10, BCNN11, VP12]. Subsampling is widely used in the first order methods, but is not as well studied for approximating the scaling matrix. In particular, theoretical guarantees are still missing.

A key challenge is that the subsampled Hessian is close to the actual Hessian along the directions corresponding to large eigenvalues (large curvature directions in $f(\beta)$), but is a poor approximation in the directions corresponding to small eigenvalues (flatter directions in $f(\beta)$). In order to overcome this problem, we use low-rank approximation. More precisely, we treat all the eigenvalues below the r-th as if they were equal to the $(r+1)$-th. This yields the desired stability with respect to the subsample. We call our algorithm NewSamp which is short for Newton-Sampling.

The rest of the chapter is organized as follows: Section 4.1.1 surveys the related work. In Section 4.2, we describe the proposed algorithm and provide the intuition behind it. Next, we present our theoretical results in Section 4.3, i.e., convergence rates corresponding to different subsampling schemes, followed by a discussion on how to choose the algorithm parameters. Two applications of the algorithm are discussed in Section 4.4. We compare our algorithm with several existing methods on various datasets in Section 4.5. Finally, in Section 4.7, we conclude with a brief discussion.

4.1.1 Related Work

Even a synthetic review of optimization algorithms for large-scale machine learning would go beyond the page limits of this chapter. Here, we emphasize that the method of choice

depends crucially on the amount of data to be used, and their dimensionality (i.e., respectively, on the parameters n and p). In this chapter, we focus on a regime in which p is large but not so large as to make matrix manipulations (of order p^2 to p^3) impossible. Also n is large but not so large as to make batch gradient computation (of order np) prohibitive. On the other hand, our aim is to avoid $\mathcal{O}(np^2)$ calculations required by standard Newton method. Examples of this regime are given in Section 4.4.

In contrast, online algorithms are the option of choice for very large n since the computation per update is independent of n. In the case of *Stochastic Gradient Descent* (SGD), the descent direction is formed by a randomly selected gradient [RM51]. Improvements to SGD have been developed by incorporating the previous gradient directions in the current update [SLRB, SHY$^+$13, Bot10, DHS11].

Batch algorithms, on the other hand, can achieve faster convergence and exploit second order information. They are competitive for intermediate n. Several methods in this category aim at quadratic, or at least super-linear convergence rates. In particular, Quasi-Newton methods have proven effective [Bis95, Nes13]. Another approach towards the same goal is to utilize subsampling to form an approximate Hessian [Mar10, BCNN11, VP12, EM15, Erd15]. If the subsampled Hessian is close to the true Hessian, these methods can approach NM in terms of convergence rate, nevertheless, they enjoy much smaller complexity per update. No convergence rate analysis is available for these methods: this analysis is the main contribution of our chapter. To the best of our knowledge, the best result in this direction is proven in [BCNN11] that estabilishes asymptotic convergence without quantitative bounds (exploiting general theory from [GNS09]).

Further improvements have been suggested either by utilizing *Conjugate Gradient* (CG) methods and/or using Krylov subspaces [Mar10, BCNN11, VP12]. Subsampling can be also used to obtain an approximate solution, if an exact solution is not required [DLFU13]. Lastly, there are various hybrid algorithms that combine two or more techniques to gain improvement. Examples include, subsampling and Quasi-Newton [SYG07, SDPG13, BHNS14], SGD and GD [FS12], NGD and NM [LRF10], NGD and low-rank approximation [LRMB08].

Algorithm 4 NewSamp

Input: $\hat{\beta}^0, r, \epsilon, \{\gamma_t, |S_t|\}_t, t = 0$.

1. **Define:** $\mathcal{P}_{\mathcal{C}}(\beta) = \text{argmin}_{\beta' \in \mathcal{C}} \|\beta - \beta'\|_2$ is the Euclidean projection onto $\mathcal{C}$, $[\mathbf{U}_k, \mathbf{\Lambda}_k] = \text{TruncatedSVD}_k(\mathbf{H})$ is the rank-k truncated SVD of $\mathbf{H}$ with $(\mathbf{\Lambda}_k)_{ii} = \lambda_i$.
2. **while** $\|\hat{\beta}^{t+1} - \hat{\beta}^t\|_2 > \epsilon$ **do**
   Subsample a set of indices $S_t \subset [n]$.
   $\mathbf{H}_{S_t} = \frac{1}{|S_t|} \sum_{i \in S_t} \boldsymbol{\nabla}^2_\beta f_i(\hat{\beta}^t)$, and $[\mathbf{U}_{r+1}, \mathbf{\Lambda}_{r+1}] = \text{TruncatedSVD}_{r+1}(\mathbf{H}_{S_t})$,
   $\mathbf{Q}^t = \lambda_{r+1}^{-1} \mathbf{I}_p + \mathbf{U}_r \left(\mathbf{\Lambda}_r^{-1} - \lambda_{r+1}^{-1} \mathbf{I}_r \right) \mathbf{U}_r^T$,
   $\hat{\beta}^{t+1} = \mathcal{P}_{\mathcal{C}} \left(\hat{\beta}^t - \gamma_t \mathbf{Q}^t \nabla_\beta f(\hat{\beta}^t) \right)$,
   $t \leftarrow t + 1$.
3. **end while**

Output: $\hat{\beta}^t$.

4.2 NewSamp: A Newton method via subsampling and eigenvalue thresholding

In the regime we consider, $n \gg p \gg 1$, there are two main drawbacks associated with the classical second order methods such as Newton method. The predominant issue in this regime is the computation of the Hessian matrix, which requires $\mathcal{O}(np^2)$ operations, and the other issue is finding the inverse of the Hessian, which requires $\mathcal{O}(p^3)$ computation. Subsampling is an effective and efficient way of addressing the first issue, by forming an approximate Hessian to exploit curvature information. Recent empirical studies show that subsampling the Hessian provides significant improvement in terms of computational cost, yet preserves the fast convergence rate of second order methods [Mar10, VP12, Erd16]. If a uniform subsample is used, the subsampled Hessian will be a random matrix with expected value at the true Hessian, which can be considered as a sample estimator to the mean. Recent advances in statistics have shown that the performance of various estimators can be significantly improved by simple procedures such as *shrinkage* and/or *thresholding* [CCS10, DGJ13, GD14, GD14]. To this extent, we use a specialized low-rank approximation as the important second order information is generally contained in the largest few eigenvalues/vectors of the Hessian. We will see in Section 4.3, how this procedure provides faster convergence rates compared to the bare subsampling methods.

NewSamp is presented as Algorithm 4. At iteration step t, the subsampled set of indices, its size and the corresponding subsampled Hessian is denoted by S_t, $|S_t|$ and $\mathbf{H}_{S_t}$, respectively. Assuming that the functions f_i's are convex, eigenvalues of the symmetric matrix $\mathbf{H}_{S_t}$ are non-negative. Therefore, singular value (SVD) and eigenvalue decompositions coincide. The operation $\text{TruncatedSVD}_k(\mathbf{H}_{S_t}) = [\mathbf{U}_k, \mathbf{\Lambda}_k]$ is the best rank-k approximation, i.e., takes $\mathbf{H}_{S_t}$ as input and returns the largest k eigenvalues in the diagonal matrix $\mathbf{\Lambda}_k \in \mathbb{R}^{k\times k}$ with the corresponding k eigenvectors $\mathbf{U}_k \in \mathbb{R}^{p\times k}$. This procedure requires $\mathcal{O}(kp^2)$ computation using a standard method, though there are faster randomized algorithms which provide accurate approximations to the truncated SVD problem with much less computational cost [HMT11]. To construct the curvature matrix $[\mathbf{Q}^t]^{-1}$, instead of using the basic rank-r approximation, we fill its 0 eigenvalues with the $(r+1)$-th eigenvalue of the subsampled Hessian which is the largest eigenvalue below the threshold. If we compute a truncated SVD with $k = r+1$ and $(\mathbf{\Lambda}_k)_{ii} = \lambda_i$, the described operation can be formulated as the following,

$$\mathbf{Q}^t = \lambda_{r+1}^{-1}\mathbf{I}_p + \mathbf{U}_r\left(\mathbf{\Lambda}_r^{-1} - \lambda_{r+1}^{-1}\mathbf{I}_r\right)\mathbf{U}_r^T, \tag{4.3}$$

which is simply the sum of a scaled identity matrix and a rank-r matrix. Note that the low-rank approximation that is suggested to improve the curvature estimation has been further utilized to reduce the cost of computing the inverse matrix. Final per-iteration cost of NewSamp will be $\mathcal{O}\left(np + (|S_t| + r)p^2\right) \approx \mathcal{O}\left(np + |S_t|p^2\right)$. NewSamp takes the parameters $\{\gamma_t, |S_t|\}_t$ and r as inputs. We discuss in Section 4.3.4, how to choose these parameters near-optimally, based on the theory we develop in Section 4.3.

Operator $\mathcal{P}_\mathcal{C}$ projects the current iterate to the feasible set $\mathcal{C}$ using Euclidean projection. Throughout, we assume that this projection can be done efficiently. In general, most unconstrained optimization problems do not require this step, and can be omitted. The purpose of projected iterations in our algorithm is mostly theoretical, and will be clear in Section 4.3.

By the construction of $\mathbf{Q}^t$, NewSamp will always be a descent algorithm. It enjoys a quadratic convergence rate at start which transitions into a linear rate in the neighborhood of the minimizer. This behavior can be observed in Figure 4.1. The left plot in Figure 1 shows the convergence behavior of NewSamp over different subsample sizes. We observe that large subsamples result in better convergence rates as expected. As the subsample size

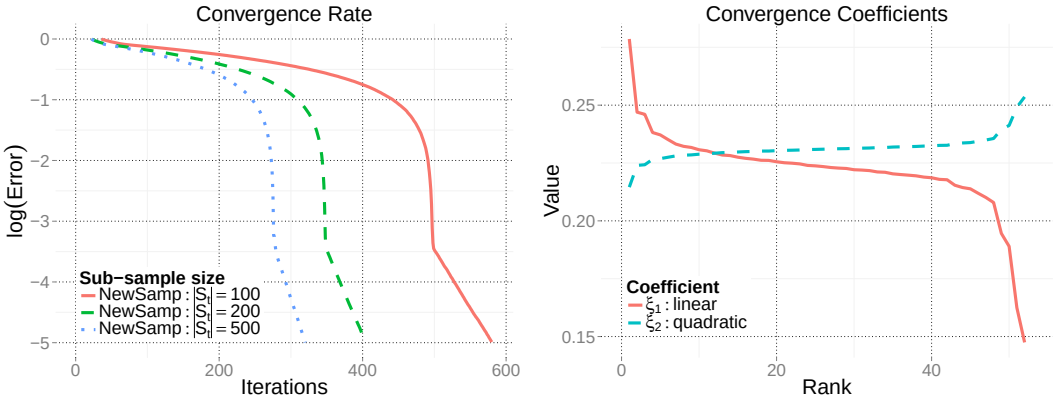

Figure 4.1: Left plot demonstrates convergence rate of NewSamp , which starts with a quadratic rate and transitions into linear convergence near the true minimizer. The right plot shows the effect of eigenvalue thresholding on the convergence coefficients. x-axis shows the number of kept eigenvalues. Plots are obtained using *Covertype* dataset.

increases, slope of the linear phase decreases, getting closer to that of quadratic phase at the transition point. This phenomenon will be explained in detail in Section 4.3, by Theorems 4.3.2 and 4.3.4. The right plot in Figure 4.1 demonstrates how the coefficients of linear and quadratic phases depend on the thresholded rank. Note that the coefficient of the quadratic phase increases with the rank threshold, whereas for the linear phase, relation is reversed.

4.3 Theoretical results

In this section, we provide the convergence analysis of NewSamp based on two different subsampling schemes:

S1: **Independent subsampling**: At each iteration t, S_t is uniformly sampled from $[n] = \{1, 2, ..., n\}$, independently from the sets $\{S_\tau\}_{\tau<t}$, with or without replacement.

S2: **Sequentially dependent subsampling**: At each iteration t, S_t is sampled from $[n]$, based on a distribution which might depend on the previous sets $\{S_\tau\}_{\tau<t}$, but not on any randomness in the data.

The first subsampling scheme is simple and commonly used in optimization. One drawback is that the subsampled set at the current iteration is independent of the previous subsamples, hence does not consider which of the samples were previously used to form

the approximate curvature information. In order to prevent cycles and obtain better performance near the optimum, one might want to increase the sample size as the iteration advances [Mar10], including previously unused samples. This process results in a sequence of dependent subsamples which falls into the subsampling scheme S2. In our theoretical analysis, we make the following assumptions:

Assumption 1 (Lipschitz continuity)**.** *For any subset $S \subset [n]$, there exists a constant $M_{|S|}$ depending on the size of S, such that $\forall \beta, \beta' \in \mathcal{C}$,*

$$\left\|\mathbf{H}_S(\beta) - \mathbf{H}_S(\beta')\right\|_2 \leq M_{|S|} \ \|\beta - \beta'\|_2.$$

Assumption 2 (Bounded Hessian)**.** *$\forall i = 1, 2, ..., n$, the Hessian of the function $f_i(\beta)$, $\boldsymbol{\nabla}^2_\beta f_i(\beta)$, is upper bounded by an absolute constant K, i.e.,*

$$\max_{i \leq n} \left\|\boldsymbol{\nabla}^2_\beta f_i(\beta)\right\|_2 \leq K.$$

4.3.1 Independent subsampling

In this section, we assume that $S_t \subset [n]$ is sampled according to the subsampling scheme S1. In fact, many stochastic algorithms assume that S_t is a uniform subset of $[n]$, because in this case the subsampled Hessian is an unbiased estimator of the full Hessian. That is, $\forall \beta \in \mathcal{C}$, $\mathbb{E}\left[\mathbf{H}_{S_t}(\beta)\right] = \mathbf{H}_{[n]}(\beta)$, where the expectation is over the randomness in S_t. We next show that for any scaling matrix $\mathbf{Q}^t$ that is formed by the subsamples S_t, iterations of the form Equation (4.2) will have a composite convergence rate, i.e., combination of a linear and a quadratic phases.

Lemma 4.3.1. *Assume that the parameter set $\mathcal{C}$ is convex and $S_t \subset [n]$ is based on subsampling scheme S1. Further, let the Assumptions 1 and 2 hold and $\beta_* \in \mathcal{C}$. Then, for an absolute constant $c > 0$, with probability at least $1 - 2/p$, the updates of the form Equation* (4.2) *satisfy*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2 + \xi_2^t \|\hat{\beta}^t - \beta_*\|_2^2,$$

for coefficients ξ_1^t *and* ξ_2^t *defined as*

$$\xi_1^t = \left\| I - \gamma_t \mathbf{Q}^t \mathbf{H}_{S_t}(\hat{\beta}^t) \right\|_2 + \gamma_t cK \left\| \mathbf{Q}^t \right\|_2 \sqrt{\frac{\log(p)}{|S_t|}},$$
$$\xi_2^t = \gamma_t \frac{M_n}{2} \left\| \mathbf{Q}^t \right\|_2 .$$

Remark 1. *If the initial point* $\hat{\beta}^0$ *is close to* β_**, the algorithm will start with a quadratic rate of convergence which will transform into linear rate later in the close neighborhood of the optimum.*

The above lemma holds for any matrix $\mathbf{Q}^t$. In particular, if we choose $\mathbf{Q}^t = \mathbf{H}_{S_t}^{-1}$, we obtain a bound for the simple subsampled Hessian method. In this case, the coefficients ξ_1^t and ξ_2^t depend on $\|\mathbf{Q}^t\|_2 = 1/\lambda_p^t$ where λ_p^t is the smallest eigenvalue of the subsampled Hessian. Note that λ_p^t can be arbitrarily small which might blow up both of the coefficients. In the following, we will see how NewSamp remedies this issue.

Theorem 4.3.2. *Let the assumptions in Lemma 4.3.1 hold. Denote by* λ_i^t*, the i-th eigenvalue of* $\mathbf{H}_{S_t}(\hat{\beta}^t)$ *where* $\hat{\beta}^t$ *is given by NewSamp at iteration step t. If the step size satisfies*

$$\gamma_t \leq \frac{2}{1 + \lambda_p^t / \lambda_{r+1}^t}, \tag{4.4}$$

then we have, with probability at least $1 - 2/p$*,*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2 + \xi_2^t \|\hat{\beta}^t - \beta_*\|_2^2,$$

for an absolute constant $c > 0$*, for the coefficients* ξ_1^t *and* ξ_2^t *are defined as*

$$\xi_1^t = 1 - \gamma_t \frac{\lambda_p^t}{\lambda_{r+1}^t} + \gamma_t \frac{cK}{\lambda_{r+1}^t} \sqrt{\frac{\log(p)}{|S_t|}},$$
$$\xi_2^t = \gamma_t \frac{M_n}{2\lambda_{r+1}^t}.$$

NewSamp has a composite convergence rate where ξ_1^t and ξ_2^t are the coefficients of the linear and the quadratic terms, respectively (See the right plot in Figure 4.1). We observe that the subsampling size has a significant effect on the linear term, whereas the quadratic term is governed by the Lipschitz constant. We emphasize that the case $\gamma_t = 1$ is feasible

for the conditions of Theorem 4.3.2. In the case of quadratic functions, since the Lipschitz constant is 0 , we obtain $\xi_2^t = 0$ and the algorithm converges linearly. Following corollary summarizes this case.

Corollary 4.3.3 (Quadratic functions)**.** *Let the assumptions of Theorem 4.3.2 hold. Further, assume that $\forall i \in [n]$, the functions $\beta : \mathbb{R}^p \to f_i(\beta)$ are quadratic. Then, for $\hat{\beta}^t$ given by NewSamp at iteration step t, for the coefficient ξ_1^t defined as in Theorem 4.3.2, with probability at least $1 - 2/p$, we have*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2. \tag{4.5}$$

4.3.2 Sequentially dependent subsampling

Here, we assume that the subsampling scheme S2 is used to generate $\{S_\tau\}_{\tau \geq 1}$. Distribution of subsampled sets may depend on each other, but not on any randomness in the dataset. Examples include fixed subsamples as well as subsamples of increasing size, sequentially covering unused data. In addition to Assumptions 1-2, we assume the following.

Assumption 3 (i.i.d. observations)**.** *Let $z_1, z_2, ..., z_n \in \mathrm{Z}$ be i.i.d. observations from a distribution* **D**. *For a fixed $\beta \in \mathbb{R}^p$ and $\forall i \in [n]$, we assume that the functions $\{f_i\}_{i=1}^n$ satisfy $f_i(\beta) = \varphi(z_i, \beta)$, for some function $\varphi : \mathrm{Z} \times \mathbb{R}^p \to \mathbb{R}$.*

Most statistical learning algorithms can be formulated as above, e.g., in classification problems, one has access to i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ where y_i and x_i denote the class label and the covariate, and φ measures the classification error (See Section 4.4 for examples). For the subsampling scheme S2, an analogue of Lemma 4.3.1 is stated in Appendix as Lemma 4.6.1, which immediately leads to the following theorem.

Theorem 4.3.4. *Assume that the parameter set $\mathcal{C}$ is convex and $S_t \subset [n]$ is based on the subsampling scheme S2. Further, let the Assumptions 1, 2 and 3 hold, almost surely. Conditioned on the event $\mathcal{E} = \{\beta_* \in \mathcal{C}\}$, if the step size satisfies Equation (4.4), then for $\hat{\beta}^t$ given by NewSamp at iteration t, with probability at least $1 - c_{\mathcal{E}}\, e^{-p}$ for $c_{\mathcal{E}} = c/\mathbb{P}(\mathcal{E})$, we have*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2 + \xi_2^t \|\hat{\beta}^t - \beta_*\|_2^2,$$

for the coefficients ξ_1^t *and* ξ_2^t *defined as*

$$\xi_1^t = 1 - \gamma_t \frac{\lambda_p^t}{\lambda_{r+1}^t} + \gamma_t \frac{c' K}{\lambda_{r+1}^t} \sqrt{\frac{p}{|S_t|} \log\left(\frac{\operatorname{diam}(\mathcal{C})^2 \left(M_n + M_{|S_t|}\right)^2 |S_t|}{K^2} \right)},$$
$$\xi_2^t = \gamma_t \frac{M_n}{2\lambda_{r+1}^t},$$

where $c, c' > 0$ *are absolute constants and* λ_i^t *denotes the* i*-th eigenvalue of* $\mathbf{H}_{S_t}(\hat{\beta}^t)$.

Compared to the Theorem 4.3.2, we observe that the coefficient of the quadratic term does not change. This is due to Assumption 1. However, the bound on the linear term is worse, since we use the uniform bound over the convex parameter set $\mathcal{C}$. The same order of magnitude is also observed by [Erd16], which relies on a similar proof technique. Similar to Corollary 4.3.3, we have the following result for the quadratic functions.

Corollary 4.3.5 (Quadratic functions)**.** *Let the assumptions of Theorem 4.3.4 hold. Further assume that* $\forall i \in [n]$*, the functions* $\beta \to f_i(\beta)$ *are quadratic. Then, conditioned on the event* $\mathcal{E}$*, with probability at least* $1 - c_{\mathcal{E}}\, e^{-p}$*, NewSamp iterates satisfy*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2,$$

for coefficient ξ_1^t *defined as in Theorem 4.3.4.*

4.3.3 Dependence of coefficients on t and convergence guarantees

The coefficients ξ_1^t and ξ_2^t depend on the iteration step which is an undesirable aspect of the above results. However, these constants can be well approximated by their analogues ξ_1^* and ξ_2^* evaluated at the optimum which are defined by simply replacing λ_j^t with λ_j^* in their definition, where the latter is the j-th eigenvalue of full-Hessian at θ_*. For the sake of simplicity, we only consider the case where the functions $\beta \to f_i(\beta)$ are quadratic.

Theorem 4.3.6. *Assume that the functions* $f_i(\beta)$ *are quadratic,* S_t *is based on scheme S1 and* $\gamma_t = 1$*. Let the full Hessian at* β_* *be lower bounded by a constant* k*. Then for sufficiently large* $|S_t|$*, we have, with probability* $1 - 2/p$

$$\left|\xi_1^t - \xi_1^*\right| \leq \frac{c_1 K \sqrt{\log(p)/|S_t|}}{k\left(k - c_2 K \sqrt{\log(p)/|S_t|}\right)} := \delta,$$

for some absolute constants c_1, c_2.

Theorem 4.3.6 implies that, when the subsampling size is sufficiently large, ξ_1^t will concentrate around ξ_1^*. Generalizing the above theorem to non-quadratic functions is straightforward, in which case, one would get additional terms involving the difference $\|\hat{\beta}^t - \beta_*\|_2$. In the case of scheme S2, if one uses fixed subsamples, i.e., $\forall t$, $S_t = S$, then the coefficient ξ_1^t does not depend on t. The following corollary gives a sufficient condition for convergence. A detailed discussion on the number of iterations until convergence and further local convergence properties can be found in Appendix C.1.

Corollary 4.3.7. *Assume that* ξ_1^t *and* ξ_2^t *are well-approximated by* ξ_1^* *and* ξ_2^* *with an error bound of* δ*, i.e.,* $\xi_i^t \leq \xi_i^* + \delta$ *for* $i = 1, 2$*, as in Theorem 4.3.6. For the initial point* $\hat{\beta}^0$*, a sufficient condition for convergence is*

$$\|\hat{\beta}^0 - \beta_*\|_2 < \frac{1 - \xi_1^* - \delta}{\xi_2^* + \delta}.$$

4.3.4 Choosing the algorithm parameters

Algorithm parameters play a crucial role in most optimization methods. Based on the theoretical results from previous sections, we discuss procedures to choose the optimal values for the step size γ_t, subsample size $|S_t|$ and rank threshold.

- *Step size:* For the step size of NewSamp at iteration t, we suggest

$$\gamma_t(\gamma) = \frac{2}{1 + \lambda_p^t/\lambda_{r+1}^t + \gamma}. \tag{4.6}$$

  where $\gamma = \mathcal{O}(\log(p)/|S_t|)$. Note that $\gamma_t(0)$ is the upper bound in Theorems 4.3.2 and 4.3.4 and it minimizes the first component of ξ_1^t. The other terms in ξ_1^t and ξ_2^t linearly depend on γ_t. To compensate for that, we shrink $\gamma_t(0)$ towards 1. Contrary to most algorithms, optimal step size of NewSamp is larger than 1. See Appendix C.2 for a rigorous derivation of Equation (4.6).

- *Sample size:* By Theorem 4.3.2, a subsample of size $\mathcal{O}((K/\lambda_p^*)^2 \log(p))$ should be sufficient to obtain a small coefficient for the linear phase. Also note that subsample size $|S_t|$ scales quadratically with the condition number.

- *Rank threshold:* For a full-Hessian with effective rank R (trace divided by the largest eigenvalue), it suffices to use $\mathcal{O}(R\log(p))$ samples [Ver10, Ver12]. Effective rank is upper bounded by the dimension p. Hence, one can use $p\log(p)$ samples to approximate the full-Hessian and choose a rank threshold which retains the important curvature information.

4.4 Examples

4.4.1 Generalized Linear Models

Finding the maximum likelihood estimator in Generalized Linear Models (GLMs) is equivalent to minimizing the negative log-likelihood $f(\beta)$,

$$\underset{\beta}{\text{minimize}}\ f(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\Phi(\langle x_i,\beta\rangle) - y_i\langle x_i,\beta\rangle\right], \tag{4.7}$$

where Φ is the *cumulant generating function*, $y_i \in \mathbb{R}$ denotes the observations, $x_i \in \mathbb{R}^p$ denotes the rows of design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $\beta \in \mathbb{R}^p$ is the coefficient vector. Note that this formulation only considers GLMs with canonical links. Here, $\langle x,\beta\rangle$ denotes the inner product between the vectors x, β. The function Φ defines the type of GLM. Well known examples include ordinary least squares (OLS) with $\Phi(z) = z^2$, logistic regression (LR) with $\Phi(z) = \log(1+e^z)$, and Poisson regression (PR) with $\Phi(z) = e^z$.

The gradient and the Hessian of the above function can be written as:

$$\nabla_\beta f(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{(1)}(\langle x_i,\beta\rangle)x_i - y_i x_i\right], \tag{4.8}$$

$$\boldsymbol{\nabla}^2_\beta f(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Psi^{(2)}(\langle x_i,\beta\rangle)x_i x_i^T. \tag{4.9}$$

We note that the Hessian of the GLM problem is always positive definite. This is because the second derivative of the cumulant generating function is simply the variance of the observations. Using the results from Section 4.3, we perform a convergence analysis of our algorithm on a GLM problem.

Corollary 4.4.1. *Let $S_t \subset [n]$ be a uniform subsample, and $\mathcal{C}$ be a convex parameter set. Assume that the second derivative of the cumulant generating function, $\Psi^{(2)}$ is bounded*

by 1*, and it is Lipschitz continuous with Lipschitz constant* L*. Further, assume that the covariates are contained in a ball of radius* $\sqrt{R_x}$*, i.e.* $\max_{i\in[n]} \|x_i\|_2 \leq \sqrt{R_x}$*. Then, for* $\hat{\beta}^t$ *given by NewSamp with constant step size* $\gamma_t = 1$ *at iteration* t*, with probability at least* $1 - 2/p$*, we have*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2 + \xi_2^t \|\hat{\beta}^t - \beta_*\|_2^2,$$

for constants ξ_1^t *and* ξ_2^t *defined as*

$$\begin{aligned}\xi_1^t =& 1 - \frac{\lambda_i^t}{\lambda_{r+1}^t} + \frac{cR_x}{\lambda_{r+1}^t}\sqrt{\frac{\log(p)}{|S_t|}},\\ \xi_2^t =& \frac{LR_x^{3/2}}{2\lambda_{r+1}^t},\end{aligned}$$

where $c > 0$ *is an absolute constant and* λ_i^t *is the* i*th eigenvalue of* $\mathbf{H}_{S_t}(\hat{\beta}^t)$*.*

Proof of Corollary 4.4.1 can be found in Appendix 4.6. Note that the bound on the second derivative is quite loose for Poisson regression due to exponentially fast growing cumulant generating function.

4.4.2 Support Vector Machines

A linear Support Vector Machine (SVM) provides a *separating hyperplane* which maximizes the *margin*, i.e., the distance between the hyperplane and the support vectors. Although the vast majority of the literature focuses on the dual problem [Vap98, SS02], SVMs can be trained using the primal as well. Since the dual problem does not scale well with the number of data points (some approaches get $\mathcal{O}(n^3)$ complexity, [WG11]), the primal might be better-suited for optimization of linear SVMs [KD05, Cha07].

The primal problem for the linear SVM can be written as

$$\underset{\beta\in\mathcal{C}}{\text{minimize}}\ f(\beta) = \frac{1}{2}\|\beta\|_2^2 + \frac{1}{2}C\sum_{i=1}^{n}\ell(y_i, \langle\beta, x_i\rangle) \tag{4.10}$$

where (y_i, x_i) denote the data samples, β defines the separating hyperplane, $C > 0$ and ℓ could be any loss function. The most commonly used loss functions include *Hinge-p loss*, *Huber loss* and their smoothed versions [Cha07]. Smoothing or approximating such losses

with more stable functions is sometimes crucial in optimization. In the case of NewSamp which requires the loss function to be twice differentiable (almost everywhere), we suggest either smoothed Huber loss, i.e.,

$$\ell(y, \langle \beta, x \rangle) = \begin{cases} 0, & \text{if } y\langle \beta, x \rangle > 3/2, \\ \frac{(3/2 - y\langle \beta, x \rangle)^2}{2}, & \text{if } |1 - y\langle \beta, x \rangle| \leq 1/2, \\ 1 - y\langle \beta, x \rangle, & \text{otherwise.} \end{cases}$$

or Hinge-2 loss, i.e.,

$$\ell(y, \langle \beta, x \rangle) = \max\{0, 1 - y\langle \beta, x \rangle\}^2 .$$

For the sake of simplicity, we will focus on Hinge-2 loss. Denote by SV_t, the set of indices of all the support vectors at iteration t, i.e.,

$$SV_t = \{i : y_i \langle \beta^t, x_i \rangle < 1\}.$$

When the loss is set to be the Hinge-2 loss, the Hessian of the SVM problem, normalized by the number of support vectors, can be written as

$$\nabla^2_\beta f(\beta) = \frac{1}{|SV_t|}\Big\{\mathbf{I} + C \sum_{i \in SV_t} x_i x_i^T\Big\}.$$

When $|SV_t|$ is large, the problem falls into our setup and can be solved efficiently using NewSamp . Note that unlike the GLM setting, Lipschitz condition of our Theorems do not apply here. However, we empirically demonstrate that NewSamp works regardless of such assumptions.

4.5 Experiments

In this section, we validate the performance of NewSamp through extensive numerical studies. We experimented on two optimization problems, namely, *Logistic Regression* (LR) and *Support Vector Machines* (SVM) with quadratic loss. LR minimizes Equation (4.7) for the logistic function, whereas SVM minimizes Equation (4.10) for the Hinge-2 loss.

For the convenience of the reader, we briefly describe the algorithms that are used in

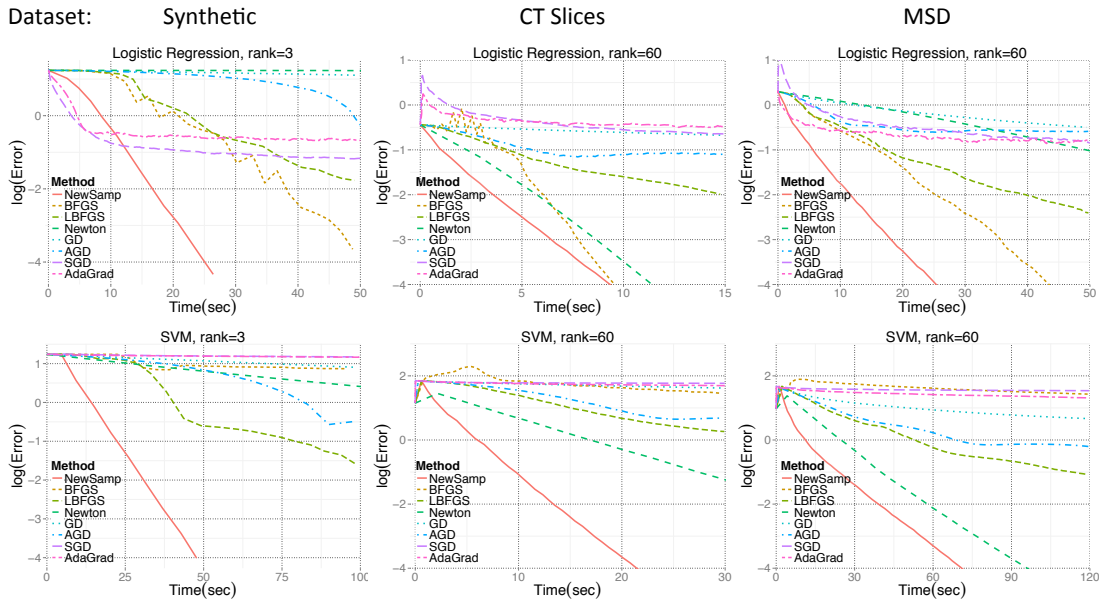

Figure 4.2: Performance of various optimization methods on different datasets. NewSamp is represented with red color .

the experiments below:

1. *Gradient Descent* (GD), at each iteration, takes a step proportional to negative of the full gradient evaluated at the current iterate. Under certain regularity conditions, GD exhibits a linear convergence rate.

2. *Accelerated Gradient Descent* (AGD) is proposed by Nesterov [Nes83], which improves over the gradient descent by using a momentum term. Performance of AGD strongly depends of the smoothness of the function f and decreasing step size adjustments may be necessary for convergence.

3. *Newton Method* (NM) achieves a quadratic convergence rate by utilizing the inverse Hessian evaluated at the current iterate. However, the computation of Hessian makes it impractical for large-scale datasets.

4. *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) is the most popular and stable Quasi-Newton method. Scaling matrix is formed by accumulating the information from iterates and gradients, satisfying *Quasi-Newton rule*. The convergence rate is locally super-linear and per-iteration cost is comparable to first order methods.

5. *Limited Memory BFGS* (L-BFGS) is a variant of BFGS, which uses only the recent iterates and gradients to form the approximate Hessian, providing significant improvement in terms of memory usage.

6. *Stochastic Gradient Descent* (SGD) is a simplified version of GD where, at each iteration, instead of the full gradient, a randomly selected gradient is used. Per-iteration cost is independent of n, yet the convergence rate is significantly slower compared to batch algorithms. We follow the guidelines of [Bot10, SHY$^+$13] for the step size,, i.e.,

$$\gamma_t = \frac{\gamma}{1 + t/c},$$

for constants $\gamma, c > 0$.

7. *Adaptive Gradient Scaling* (AdaGrad) is an online algorithm which uses an adaptive learning rate based on the previous gradients. AdaGrad significantly improves the performance and stability of SGD [DHS11]. This is achieved by scaling each entry of gradient differently. , i.e., at iteration step t, step size for the j-th coordinate is

$$(\gamma_t)_j = \frac{\gamma}{\sqrt{\delta + \sum_{\tau=1}^{t} (\nabla_\beta f(\hat{\beta}^t))_j}},$$

for constants $\delta, \gamma > 0$.

For each of the batch algorithms, we used constant step size, and for all the algorithms, we choose the step size that provides the fastest convergence. For the stochastic algorithms, we optimized over the parameters that define the step size. Parameters of NewSamp are selected following the guidelines described in Section 4.3.4.

We experimented over various datasets that are given in Table 4.1. The real datasets are downloaded from the UCI repository [Lic13]. Each dataset consists of a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the corresponding observations (classes) $y \in \mathbb{R}^n$. Synthetic data is generated through a multivariate Gaussian distribution with a randomly generated covariance matrix. As a methodological choice, we selected moderate values of p, for which Newton Method can still be implemented, and nevertheless we can demonstrate an improvement. For larger values of p, comparison is even more favorable to our approach.

The effects of subsampling size $|S_t|$ and rank threshold are demonstrated in Figure 4.1. A thorough comparison of the aforementioned optimization techniques is presented in

Figure 4.2. In the case of LR, we observe that stochastic algorithms enjoy fast convergence at start, but slows down later as they get close to the true minimizer. The algorithm that comes close to NewSamp in terms of performance is BFGS. In the case of SVM, Newton method is the closest algorithm to NewSamp, yet in all scenarios, NewSamp outperforms its competitors. Note that the global convergence of BFGS is not better than that of GD [Nes13]. The condition for super-linear rate is $\sum_t \|\beta^t - \beta_*\|_2 < \infty$ for which, an initial point close to the optimum is required [DM77]. This condition can be rarely satisfied in practice, which also affects the performance of the other second order methods. For NewSamp , even though the rank thresholding provides a certain level of robustness, we observed that the choice of a good starting point is still an important factor. Details about Figure 4.2 can be found in Table C.2 in Appendix. For additional experiments and a detailed discussion, see Appendix C.3.

Dataset	n	p	r	Reference
CT slices	53500	386	60	[GKS$^+$11]
Covertype	581012	54	20	[BD99]
MSD	515345	90	60	[MEWL11]
Synthetic	500000	300	3	-

Table 4.1: Datasets used in the experiments.

4.6 Proof of Main Results

4.6.1 Proofs of Lemma 4.3.1 and Theorem 4.3.2

Proof of Lemma 4.3.1. We write,

$$\begin{aligned}\hat{\beta}^t - \beta_* - \gamma_t \mathbf{Q}^t \nabla_\beta f(\hat{\beta}^t) &= \hat{\beta}^t - \beta_* - \gamma_t \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta f(\beta_* + \tau(\hat{\beta}^t - \beta_*))(\hat{\beta}^t - \beta_*)\, d\tau, \\ &= \left(I - \gamma_t \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta f(\beta_* + \tau(\hat{\beta}^t - \beta_*)) d\tau \right) (\hat{\beta}^t - \beta_*)\,.\end{aligned}$$

Since the projection $\mathcal{P}_{\mathcal{C}}$ in step 2 of NewSamp can only decrease the ℓ_2 distance, we obtain

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \left\| I - \gamma_t \mathbf{Q}^t \int_0^1 \nabla^2_\beta f(\beta_* + \tau(\hat{\beta}^t - \beta_*)) d\tau \right\|_2 \|\hat{\beta}^t - \beta_*\|_2.$$

Note that the first term on the right hand side governs the convergence behavior of the algorithm.

Next, for an index set $S \subset [n]$, define the matrix $\mathbf{H}_S(\beta)$ as

$$\mathbf{H}_S(\beta) = \frac{1}{|S|} \sum_{i \in S} \mathbf{H}_i(\beta)$$

where $|S|$ denotes the size of the set. Denote the integral in the above equation by $\widetilde{\mathbf{H}}$, that is,

$$\widetilde{\mathbf{H}} = \int_0^1 \nabla^2_\beta f(\beta_* + \tau(\hat{\beta}^t - \beta_*)) d\tau.$$

By the triangle inequality, the governing term that determines the convergence rate can be bounded as

$$\begin{aligned} \left\| I - \gamma_t \mathbf{Q}^t \widetilde{\mathbf{H}} \right\|_2 \leq & \left\| I - \gamma_t \mathbf{Q}^t \mathbf{H}_S(\hat{\beta}^t) \right\|_2 \\ & + \gamma_t \left\| \mathbf{Q}^t \right\|_2 \left\{ \left\| \mathbf{H}_S(\hat{\beta}^t) - \mathbf{H}_{[n]}(\hat{\beta}^t) \right\|_2 + \left\| \mathbf{H}_{[n]}(\hat{\beta}^t) - \widetilde{\mathbf{H}} \right\|_2 \right\}, \end{aligned} \tag{4.11}$$

which holds, regardless of the choice of $\mathbf{Q}^t$.

In the following, we will use some matrix concentration results to bound the right hand side of Equation (4.11). The result for sampling with replacement can be obtained by matrix Hoeffding's inequality given in [Tro12]. Note that this explicitly assumes that the samples are independent. For the concentration bounds under sampling without replacement (see i.e. [GN10, Gro11, MJC+14]), we will use the Operator-Bernstein inequality given in [GN10] which is provided in Section C.4 as Lemma C.4.2 for convenience.

Using any indexing over the elements of subsample S, we denote the each element in S by s_i, i.e.,

$$S = \{s_1, s_2, ..., s_{|S|}\}.$$

For $\beta \in \mathcal{C}$, we define the centered Hessians, $\mathcal{W}_i(\beta)$ as

$$\mathcal{W}_i(\beta) = \mathbf{H}_{s_i}(\beta) - \mathbb{E}\left[\mathbf{H}_{s_i}(\beta)\right],$$

where the $\mathbb{E}\left[\mathbf{H}_{s_i}(\beta)\right]$ is just the full Hessian at β.

By the Assumption (2), we have

$$\begin{aligned} &\max_{i\leq n} \|\mathbf{H}_i(\beta)\|_2 = \left\|\boldsymbol{\nabla}^2_\beta f_i(\beta)\right\|_2 \leq K, \\ &\max_{i\leq n} \|\mathcal{W}_i\|_2 \leq 2K \coloneqq \gamma, \qquad \max_{i\leq n} \left\|\mathcal{W}_i^2\right\|_2 \leq 4K^2 \coloneqq \sigma^2. \end{aligned} \tag{4.12}$$

Next, we apply the matrix Bernstein's inequality given in Lemma C.4.2. For $\epsilon \leq 4K$, and $\beta \in \mathcal{C}$,

$$\mathbb{P}\left(\left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 > \epsilon\right) \leq 2p \exp\left\{-\frac{\epsilon^2|S|}{16K^2}\right\}. \tag{4.13}$$

Therefore, to obtain a convergence rate of $\mathcal{O}(1/p)$, we let

$$\epsilon = C\sqrt{\frac{\log(p)}{|S|}},$$

where $C = 6K$ is sufficient. We also note that the condition on ϵ is trivially satisfied by our choice of ϵ in the target regime.

For the last term, we may write,

$$\begin{aligned} \left\|\mathbf{H}_{[n]}(\hat{\theta}^t) - \widetilde{\mathbf{H}}\right\|_2 &= \left\|\mathbf{H}_{[n]}(\hat{\theta}^t) - \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \tau(\hat{\theta}^t - \theta_*))d\tau\right\|_2, \\ &\leq \int_0^1 \left\|\mathbf{H}_{[n]}(\hat{\theta}^t) - \boldsymbol{\nabla}^2_\theta f(\theta_* + \tau(\hat{\theta}^t - \theta_*))\right\|_2 d\tau, \\ &\leq \int_0^1 M_n(1-\tau)\|\hat{\theta}^t - \theta_*\|_2 d\tau, \\ &= \frac{M_n}{2}\|\hat{\theta}^t - \theta_*\|_2. \end{aligned}$$

First inequality follows from the fact that norm of an integral is less than or equal to the integral of the norm. Second inequality follows from the Lipschitz property.

Combining the above results, we obtain the following for the governing term in Equation (4.11): For some absolute constants $c, C > 0$, with probability at least $1 - 2/p$, we have

$$\begin{aligned}&\left\|I - \gamma_t \mathbf{Q}^t \mathbf{H}_{[n]}(\tilde{\beta}^t)\right\|_2 \\ &\leq \left\|I - \gamma_t \mathbf{Q}^t \mathbf{H}_S(\hat{\beta}^t)\right\|_2 + \gamma_t \left\|\mathbf{Q}^t\right\|_2 \left\{6K\sqrt{\frac{\log(p)}{|S|}} + \frac{M_n}{2}\ \|\hat{\beta}^t - \beta_*\|_2\right\}.\end{aligned}$$

Hence, the proof is completed. □

Proof of Theorem 4.3.2. Using the definition of $\mathbf{Q}^t$ in NewSamp, we immediately obtain that

$$\left\|I - \gamma_t \mathbf{Q}^t \mathbf{H}_{S_t}(\hat{\beta}^t)\right\|_2 = \max_{i>r}\left\{\left|1 - \gamma_t \frac{\lambda_i^t}{\lambda_{r+1}^t}\right|\right\}, \tag{4.14}$$

and that $\left\|\mathbf{Q}^t\right\|_2 = 1/\lambda_{r+1}^t$. Then the proof follows from Lemma 4.3.1 and by the assumption on the step size. □

4.6.2 Proof of Theorem 4.3.6

Lemma 4.6.1. *Assume that the parameter set $\mathcal{C}$ is convex and $S_t \subset [n]$ is based on subsampling scheme S2. Further, let the Assumptions 1, 2 and 3 hold, almost surely. Then, for some absolute constants $c, C > 0$, with probability at least $1 - e^{-p}$, the updates of the form stated in Equation* (4.2) *satisfy*

$$\|\hat{\beta}^{t+1} - \beta_*\|_2 \leq \xi_1^t \|\hat{\beta}^t - \beta_*\|_2 + \xi_2^t \|\hat{\beta}^t - \beta_*\|_2^2,$$

for coefficients ξ_1^t, ξ_2^t defined as

$$\begin{aligned}\xi_1^t =& \left\|I - \gamma_t \mathbf{Q}^t \mathbf{H}_{S_t}(\hat{\beta}^t)\right\|_2 + \gamma_t \left\|\mathbf{Q}^t\right\|_2 cK\sqrt{\frac{p}{|S_t|}\log\left(\frac{\operatorname{diam}(\mathcal{C})^2\left(M_n + M_{|S_t|}\right)^2|S_t|}{K^2}\right)} \\ \xi_2^t =& \gamma_t \frac{M_n}{2}\left\|\mathbf{Q}^t\right\|_2.\end{aligned}$$

Proof of Lemma 4.6.1. The first part of the proof is the same as Lemma 4.3.1. We carry our analysis from Equation (4.11). Note that in this general set-up, the iterates are random

variables that depend on the random functions. Therefore, we use a uniform bound for the right hand side in Equation (4.11). That is,

$$\left\|I - \gamma_t \mathbf{Q}^t \widetilde{\mathbf{H}}\right\|_2 \leq \left\|I - \gamma_t \mathbf{Q}^t \mathbf{H}_S(\hat{\beta}^t)\right\|_2 + \gamma_t \left\|\mathbf{Q}^t\right\|_2 \Big\{ \sup_{\beta \in \mathcal{C}} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 + \frac{M_n}{2} \|\hat{\theta}^t - \theta_*\|_2 \Big\}.$$

By the Assumption 1, given $\beta, \beta' \in \mathcal{C}$ such that $\|\beta - \beta'\|_2 \leq \Delta$, we have,

$$\begin{aligned} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 &\leq \left\|\mathbf{H}_S(\beta') - \mathbf{H}_{[n]}(\beta')\right\|_2 + \left(M_n + M_{|S|}\right) \|\beta - \beta'\|_2 \\ &\leq \left\|\mathbf{H}_S(\beta') - \mathbf{H}_{[n]}(\beta')\right\|_2 + \left(M_n + M_{|S|}\right) \Delta. \end{aligned}$$

Next, we will use a covering net argument to obtain a bound on the matrix empirical process. Note that similar bounds on the matrix forms can be obtained through other approaches like *chaining* as well [DE17]. Let $\mathcal{T}_\Delta$ be a Δ-net over the convex set $\mathcal{C}$. By the above inequality, we obtain

$$\sup_{\beta \in \mathcal{C}} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 \leq \max_{\beta' \in \mathcal{T}_\Delta} \left\|\mathbf{H}_S(\beta') - \mathbf{H}_{[n]}(\beta')\right\|_2 + \left(M_n + M_{|S|}\right) \Delta. \tag{4.15}$$

Now we will argue that the right hand side is small with high probability using the matrix Hoeffding's inequality from [Tro12]. By the union bound over $\mathcal{T}_\Delta$, we have

$$\mathbb{P}\left(\max_{\beta' \in \mathcal{T}_\Delta} \left\|\mathbf{H}_S(\beta') - \mathbf{H}_{[n]}(\beta')\right\|_2 > \epsilon \right) \leq |\mathcal{T}_\Delta| \; \mathbb{P}\left(\left\|\mathbf{H}_S(\beta') - \mathbf{H}_{[n]}(\beta')\right\|_2 > \epsilon \right).$$

For the first term on the right hand side, by Lemma C.4.1, we write:

$$|\mathcal{T}_\Delta| \leq \left(\frac{\operatorname{diam}(\mathcal{C})}{2\Delta/\sqrt{p}} \right)^p.$$

As before, let $S = \{s_1, s_2, ..., s_{|S|}\}$, that is, s_i denote the different indices in S. For any

$\beta \in \mathcal{C}$ and $i = 1, 2, ..., n$, we define the centered Hessians $\mathcal{W}_i(\beta)$ as

$$\mathcal{W}_i(\beta) = \mathbf{H}_{s_i}(\beta) - \mathbf{H}_{[n]}(\beta).$$

By the Assumption (2), we have the same bounds as in Equation (4.12). Hence, for $\epsilon > 0$ and $\beta \in \mathcal{C}$, by the matrix Hoeffding's inequality [Tro12],

$$\mathbb{P}\left(\left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 > \epsilon\right) \leq \ 2p \exp\left\{-\frac{|S|\epsilon^2}{32K^2}\right\}.$$

We would like to obtain an exponential decay with a rate of at least $\mathcal{O}(p)$. Hence, we require,

$$\begin{aligned} p \log\left(\frac{\text{diam}(\mathcal{C})\sqrt{p}}{2\Delta}\right) + \log(2p) + p \leq \ & p \log\left(\frac{4\text{diam}(\mathcal{C})\sqrt{p}}{\Delta}\right), \\ \leq & \frac{|S|\epsilon^2}{32K^2}, \end{aligned}$$

which gives the optimal value of ϵ as

$$\epsilon \geq \sqrt{\frac{32K^2p}{|S|} \log\left(\frac{4\text{diam}(\mathcal{C})\sqrt{p}}{\Delta}\right)}.$$

Therefore, we conclude that for the above choice of ϵ, with probability at least $1 - e^{-p}$, we have

$$\max_{\beta \in \mathcal{T}_\Delta} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 < \sqrt{\frac{32K^2p}{|S|} \log\left(\frac{4\text{diam}(\mathcal{C})\sqrt{p}}{\Delta}\right)}.$$

Applying this result to the inequality in Equation (4.15), we obtain that with probability at least $1 - e^{-p}$,

$$\sup_{\beta \in \mathcal{C}} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 \leq \sqrt{\frac{32K^2p}{|S|} \log\left(\frac{4\text{diam}(\mathcal{C})\sqrt{p}}{\Delta}\right)} + \left(M_n + M_{|S|}\right)\Delta.$$

The right hand side of the above inequality depends on the net covering diameter Δ. We optimize over Δ using Lemma B.4.6 which provides for

$$\Delta = 4\sqrt{\frac{K^2 p}{\left(M_n + M_{|S|}\right)^2 |S|} \log\left(\frac{\operatorname{diam}(\mathcal{C})^2 \left(M_n + M_{|S|}\right)^2 |S|}{K^2}\right)},$$

we obtain that with probability at least $1 - e^{-p}$,

$$\sup_{\beta \in \mathcal{C}} \left\|\mathbf{H}_S(\beta) - \mathbf{H}_{[n]}(\beta)\right\|_2 \leq 8K\sqrt{\frac{p}{|S|} \log\left(\frac{\operatorname{diam}(\mathcal{C})^2 \left(M_n + M_{|S|}\right)^2 |S|}{K^2}\right)}.$$

Combining this with the bound stated in Equation (4.11), we conclude the proof. □

4.6.3 Proofs of Theorem 4.3.6 and Corollary 4.4.1

Proof of Theorem 4.3.6.

$$\begin{aligned}\left|\xi_1^t - \xi_1^*\right| =& \left|\frac{\lambda_p^t}{\lambda_{r+1}^t} - \frac{\lambda_p^*}{\lambda_{r+1}^*}\right| + cK\sqrt{\frac{\log(p)}{|S_t|}}\left|\frac{1}{\lambda_{r+1}^t} - \frac{1}{\lambda_{r+1}^*}\right| \\ \leq& \frac{K|\lambda_{r+1}^t - \lambda_{r+1}^*| + K|\lambda_p^t - \lambda_p^*|}{\lambda_{r+1}^* \lambda_{r+1}^t} + cK\sqrt{\frac{\log(p)}{|S_t|}}\frac{|\lambda_{r+1}^t - \lambda_{r+1}^*|}{\lambda_{r+1}^* \lambda_{r+1}^t}\end{aligned}$$

By the Weyl's and matrix Hoeffding's [Tro12] inequalities (See Equation (4.13) for details), we can write

$$|\lambda_j^t - \lambda_j^*| \leq \left\|\mathbf{H}_{S_t}(\hat{\beta}^t) - \mathbf{H}_{[n]}(\beta_*)\right\|_2 \leq cK\sqrt{\frac{\log(p)}{|S_t|}},$$

with probability $1 - 2/p$. Then,

$$\begin{aligned}\left|\xi_1^t - \xi_1^*\right| \leq& \frac{c' K\sqrt{\frac{\log(p)}{|S_t|}}}{\lambda_{r+1}^* \lambda_{r+1}^t} + \frac{c'' K^2 \frac{\log(p)}{|S_t|}}{\lambda_{r+1}^* \lambda_{r+1}^t}, \\ \leq& \frac{c''' K\sqrt{\frac{\log(p)}{|S_t|}}}{k\left(k - cK\sqrt{\frac{\log(p)}{|S_t|}}\right)},\end{aligned}$$

for some constants c and c'''. □

Proof of Corollary 4.4.1. Observe that $f_i(\beta) = \Phi(\langle x_i, \beta\rangle) - y_i\langle x_i, \beta\rangle$, and $\boldsymbol{\nabla}^2_\beta f_i(\beta) = x_i x_i^T \Psi^{(2)}(\langle x_i, \beta\rangle)$. For an index set S, we have $\forall \beta, \beta' \in \mathcal{C}$

$$\begin{aligned}\left\|\mathbf{H}_S(\beta) - \mathbf{H}_S(\beta')\right\|_2 &= \left\| \frac{1}{|S|} \sum_{i\in S} x_i x_i^T \left[\Psi^{(2)}(\langle x_i, \beta\rangle) - \Psi^{(2)}(\langle x_i, \beta'\rangle)\right] \right\|_2, \\ &\leq L \max_{i\in S} \|x_i\|_2^3 \ \|\beta - \beta'\|_2 \leq L R_x^{3/2} \ \|\beta - \beta'\|_2.\end{aligned}$$

Therefore, the Assumption 1 is satisfied with the Lipschitz constant $M_{|S_t|} \coloneqq L R_x^{3/2}$. Moreover, by the inequality

$$\left\|\boldsymbol{\nabla}^2_\beta f_i(\beta)\right\|_2 = \|x_i\|_2^2 \ \Psi^{(2)}(\langle x_i, \beta\rangle) \leq R_x, = \left\| x_i x_i^T \Psi^{(2)}(\langle x_i, \beta\rangle)\right\|_2$$

the Assumption 2 is satisfied for $K \coloneqq R_x$. We conclude the proof by applying Theorem 4.3.2. □

4.7 Discussion

In this chapter, we proposed a subsampling based second order method utilizing low-rank Hessian estimation. The proposed method has the target regime $n \gg p$ and has $\mathcal{O}\left(np + |S|p^2\right)$ complexity per-iteration. We showed that the convergence rate of NewSamp is composite for two widely used subsampling schemes, i.e., starts as quadratic convergence and transforms to linear convergence near the optimum. Convergence behavior under other subsampling schemes is an interesting line of research. Numerical experiments on both real and synthetic datasets demonstrate the performance of the proposed algorithm which we compared to the classical optimization methods.

Chapter 5

Conclusion

Methods and techniques that we have presented in this dissertation rely on a broad range of topics from statistics, optimization, machine learning, and applied probability. The connections among these fields have become more essential lately, mainly because of the recent advances in computational resources, the availability of large amount of data, and the consequent growing interest in statistical and machine learning algorithms. More specifically, our focus was on designing computationally efficient estimation and prediction techniques for various statistical learning problems in large-scale and/or high-dimensional settings. Using tools from statistics and probability theory such as Stein's lemma, subsampling and shrinkage techniques, we developed scalable algorithms for various data science problems, and understood their theoretical guarantees and statistical limitations.

Recent advances in computational resources introduced many challenges to modern statistical sciences. However, we have seen in this dissertation that there are many strong tools in statistics that can be used to remedy these issues. Our main tool, Stein's lemma, has been at the focus of statisticians since Charles Stein's seminal work in 1981 [Ste81]. It is fascinating that despite its simplicity, Stein's lemma has countless applications in statistical estimation, and probability theory. Yet for another application of Stein's lemma, we have seen that it can be very useful in designing optimization algorithms for large-scale problems.

In this dissertation, we have reversed the classical arrangement between statistics and optimization, and argued that optimization algorithms can also immensely benefit from the classical results from statistical estimation theory as well.

Appendix A

Supplement for Chapter 2

A.1 Auxiliary Lemmas

Lemma A.1.1 (Sub-exponential vector concentration)**.** *Let $x_1, x_2, ..., x_n$ be independent centered sub-exponential random vectors with $\max_i \|x_i\|_{\psi_1} = \kappa$. Then we have*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} x_i\right\|_2 > c\kappa\sqrt{\frac{p}{n}}\right) \leq \exp\{-p\}. \tag{A.1}$$

whenever $n > 4c^2 p$ for an absolute constant c.

Proof of Lemma A.1.1. For a vector $z \in \mathbb{R}^p$, we have $\|z\|_2 = \sup_{\|u\|_2=1} \langle u, z\rangle$ since the dual of ℓ_2 norm is itself. Therefore, we write

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} x_i\right\|_2 > t\right) = \mathbb{P}\left(\sup_{\|u\|_2=1} \frac{1}{n}\sum_{i=1}^{n} \langle u, x_i\rangle > t\right).$$

Now, let $\mathcal{N}_\epsilon$ be an ϵ-net over $\mathcal{S}^{p-1} = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$, and observe that

$$\begin{aligned}\max_{u\in\mathcal{N}_\epsilon} \langle u, x\rangle \geq& (1-\epsilon) \sup_{\|u\|_2=1} \langle u, x\rangle,\\ =& (1-\epsilon)\|x\|_2,\end{aligned}$$

with $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^p$. Hence, we may write

$$\begin{aligned}\mathbb{P}\left(\sup_{\|u\|_2=1}\frac{1}{n}\sum_{i=1}^{n}\langle u, x_i\rangle > t\right) \leq & \mathbb{P}\left(\max_{u\in\mathcal{N}_\epsilon}\frac{1}{n}\sum_{i=1}^{n}\langle u, x_i\rangle > t(1-\epsilon)\right), \\ & \leq |\mathcal{N}_\epsilon|\,\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\langle u, x_i\rangle > t(1-\epsilon)\right).\end{aligned}$$

For any $u \in \mathcal{S}^{p-1}$, we have $\|\langle u, x_i\rangle\|_{\psi_1} \leq \kappa$. Then, by the Bernstein-type inequality for sub-exponential random variables [Ver10], we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\langle u, x_i\rangle > t(1-\epsilon)\right) \leq \exp\left\{-cn\min\left\{\frac{t^2(1-\epsilon)^2}{\kappa^2}, \frac{t(1-\epsilon)}{\kappa}\right\}\right\},$$

for an absolute constant c. Therefore, the probability on the left hand side of Equation (A.1) can be bounded by

$$\left(1+\frac{2}{\epsilon}\right)^p \exp\left\{-cn\frac{t^2(1-\epsilon)^2}{\kappa^2}\right\} = \exp\left\{-cn\frac{t^2(1-\epsilon)^2}{\kappa^2} + p\log\left(1+\frac{2}{\epsilon}\right)\right\},$$

whenever $t < \kappa/(1-\epsilon)$. Choosing $\epsilon = 0.5$ and for an absolute constant $c' > 3.24/c$ and letting

$$t = c'\kappa\sqrt{\frac{p}{n}},$$

we conclude the proof. □

Lemma A.1.2. *Let $B(\tilde{\beta})$ denote the ball centered around $\tilde{\beta}$ with radius δ, i.e.,*

$$B(\tilde{\beta}) = \left\{\beta : \left\|\beta - \tilde{\beta}\right\|_2 \leq \delta\right\}.$$

For $i = 1, ..., n$, let $x_i \in \mathbb{R}^p$ be i.i.d. centered sub-Gaussian random vectors with norm bounded by κ and $\mathbb{E}\left[\|x\|_2\right] = \tilde{\mu}\sqrt{p}$. Given a function $g : \mathbb{R} \to \mathbb{R}$ that is uniformly bounded by $b > 0$, and Lipschitz continuous with k,

$$\mathbb{P}\left(\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^{n} g(\langle x_i, \beta\rangle) - \mathbb{E}\left[g(\langle x, \beta\rangle)\right]\right| > c(b+\kappa/\tilde{\mu})\sqrt{\frac{p}{n/\log(n)}}\right) \leq 2\exp\{-p\},$$

whenever $np > 51\max\{\chi, \chi^{-1}\}$ *for* $\chi = (b + \kappa/\tilde{\mu})^2/(c\delta^2 k^2 \tilde{\mu}^2)$. *Above,* c *is an absolute constant.*

Proof of Lemma A.1.2. Let $\mathbb{E}[\|x\|_2] = \mu = \tilde{\mu}\sqrt{p}$ and for $\epsilon > 0$, $\beta \in B(\tilde{\beta})$ and $w \in \mathbb{R}^p$ define the bounding functions

$$\begin{aligned} l_\beta(w) =& g(\langle w, \beta\rangle) - \epsilon\|w\|_2/4\mu, \\ u_\beta(w) =& g(\langle w, \beta\rangle) + \epsilon\|w\|_2/4\mu. \end{aligned}$$

Let $\mathcal{N}_\Delta$ be a net over $B(\tilde{\beta})$ in the sense that for any $\beta_1 \in B(\tilde{\beta})$, $\exists\beta_2 \in \mathcal{N}_\Delta$ such that $\|\beta_1 - \beta_2\|_2 \leq \Delta$. We fix $\Delta_* = \epsilon/(4k\mu)$ and write $\forall\beta_1 \in B$, $\exists\beta_2 \in \mathcal{N}_{\Delta_*}$,

1. an upper bound of the form:

$$\begin{aligned} g(\langle w, \beta_1\rangle) \leq& g(\langle w, \beta_2\rangle) + k\left|\langle w, \beta_1 - \beta_2\rangle\right|, \\ \leq& g(\langle w, \beta_2\rangle) + k\left\|w\right\|_2 \Delta_*, \\ =& u_{\beta_2}(w), \end{aligned}$$

2. and a lower bound of the form:

$$\begin{aligned} g(\langle w, \beta_1\rangle) \geq& g(\langle w, \beta_2\rangle) - k\left|\langle w, \beta_1 - \beta_2\rangle\right|, \\ \geq& g(\langle w, \beta_2\rangle) - k\left\|w\right\|_2 \Delta_*, \\ =& l_{\beta_2}(w), \end{aligned}$$

where the second steps in the above inequalities follow from the Cauchy-Schwarz inequality. These functions are called *bracketing functions* in the context of empirical process theory.

Hence, we can write that $\forall\beta_1 \in B(\tilde{\beta})$, $\exists\beta_2 \in \mathcal{N}_{\Delta_*}$ such that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n l_{\beta_2}(x_i) - \mathbb{E}\left[l_{\beta_2}(x)\right] - \epsilon/2 \leq& \frac{1}{n}\sum_{i=1}^n g(\langle x_i, \beta_1\rangle) - \mathbb{E}\left[g(\langle x, \beta_1\rangle)\right], \\ \leq& \frac{1}{n}\sum_{i=1}^n u_{\beta_2}(x_i) - \mathbb{E}\left[u_{\beta_2}(x)\right] + \epsilon/2. \end{aligned}$$

The above inequalities translate to the following conclusion: Whenever the following

event happens,

$$\left\{\left|\frac{1}{n}\sum_{i=1}^{n} g(\langle x_i, \beta_1\rangle) - \mathbb{E}\left[g(\langle x, \beta_1\rangle)\right]\right| > \epsilon\right\},$$

at least one of the following events happens

$$\left\{\frac{1}{n}\sum_{i=1}^{n} u_{\beta_2}(x_i) - \mathbb{E}\left[u_{\beta_2}(x)\right] > \epsilon/2\right\} \quad \text{or} \quad \left\{\frac{1}{n}\sum_{i=1}^{n} l_{\beta_2}(x_i) - \mathbb{E}\left[l_{\beta_2}(x)\right] < -\epsilon/2\right\}.$$

Therefore, using the union bound on the above events, we may obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\beta\in B(\tilde{\beta})}\left|\frac{1}{n}\sum_{i=1}^{n} g(\langle x_i, \beta\rangle) - \mathbb{E}\left[g(\langle x, \beta\rangle)\right]\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(\max_{\beta\in\mathcal{N}_{\Delta_*}}\frac{1}{n}\sum_{i=1}^{n} u_{\beta}(x_i) - \mathbb{E}\left[u_{\beta}(x)\right] > \epsilon/2\right) \\
&\quad + \mathbb{P}\left(\max_{\beta\in\mathcal{N}_{\Delta_*}}\frac{1}{n}\sum_{i=1}^{n} l_{\beta}(x_i) - \mathbb{E}\left[l_{\beta}(x)\right] < -\epsilon/2\right).
\end{aligned} \tag{A.2}$$

Note that the right hand side of the above inequality has two terms both of which are of the same form. For simplicity, we bound only the first one. The bound for the second one follows from the exact same steps.

The relation between sub-Gaussian and sub-exponential norms [Ver10] allows us to write

$$\|\|x\|_2\|_{\psi_2}^2 \leq \|\|x\|_2^2\|_{\psi_1} \leq \sum_{i=1}^{p}\|x_i^2\|_{\psi_1}, \leq 2\sum_{i=1}^{p}\|x_i\|_{\psi_2}^2 \leq 2\kappa^2 p, \tag{A.3}$$

where the second step follows from the triangle inequality. Hence, we conclude that $\|x\|_2 - \mathbb{E}\left[\|x\|_2\right]$ is a centered sub-Gaussian random variable with norm upper bounded by $3\kappa\sqrt{p}$.

For $\epsilon < 4/3$, we notice that the random variable $u_\beta(x) = g(\langle x, \beta\rangle) + \epsilon\|x\|_2/4\mu$ is also sub-Gaussian with norm

$$\|u_\beta(x)\|_{\psi_2} \leq b + \frac{\epsilon}{4\tilde{\mu}}3\kappa \leq b + \kappa/\tilde{\mu},$$

and consequently, the centered random variable $u_\beta(x) - \mathbb{E}\left[u_\beta(x)\right]$ has the sub-Gaussian norm upper bounded by $2b + 2\kappa/\tilde{\mu}$.

Then, by the Hoeffding-type inequality for the sub-Gaussian random variables, we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} u_\beta(x_i) - \mathbb{E}\left[u_\beta(x)\right] > \epsilon/2\right) \leq \exp\left\{-cn\frac{\epsilon^2}{(b+\kappa/\tilde{\mu})^2}\right\}$$

for an absolute constant $c > 0$.

By the same argument above, one can obtain the same result for the function $l_\beta(x)$. Using Hoeffding bounds in Equation (A.2) along with the union bound over the net, we immediately obtain

$$\mathbb{P}\left(\sup_{\beta\in B(\tilde{\beta})}\left|\frac{1}{n}\sum_{i=1}^{n} g(\langle x_i,\beta\rangle) - \mathbb{E}\left[g(\langle x,\beta\rangle)\right]\right| > \epsilon\right) \leq 2\left|\mathcal{N}_{\Delta_*}\right|\exp\left\{-cn\frac{\epsilon^2}{(b+\kappa/\tilde{\mu})^2}\right\}$$

for some absolute constant c.

Using a standard covering argument over the net $\mathcal{N}_{\Delta_*}$ as given in Lemma C.4.1, we have

$$\left|\mathcal{N}_{\Delta_*}\right| \leq \left(\frac{\delta\sqrt{p}}{\Delta_*}\right)^p = \left(\frac{4\delta k\tilde{\mu}p}{\epsilon}\right)^p.$$

Combining this with the previous bound, and choosing

$$\epsilon^2 = \frac{p}{n}\frac{(b+\kappa/\tilde{\mu})^2}{2c}\log\left(\frac{32c\delta^2k^2\tilde{\mu}^2pn}{(b+\kappa/\tilde{\mu})^2}\right)$$

we get

$$\begin{aligned}
&2\left(\frac{4\delta k\tilde{\mu}p}{\epsilon}\right)^p\exp\left\{-cn\frac{\epsilon^2}{(b+\kappa/\tilde{\mu})^2}\right\} = 2\exp\left\{-\frac{p}{2}\log\log\left(\frac{32c\delta^2k^2\tilde{\mu}^2pn}{(b+\kappa/\tilde{\mu})^2}\right)\right\}\\
&\leq 2\exp\left\{-p\right\},
\end{aligned}$$

whenever $np > 51\max\{\chi,\chi^{-1}\}$ for $\chi = (b+\kappa/\tilde{\mu})^2/(c\delta^2k^2\tilde{\mu}^2)$.

□

Lemma A.1.3 (Corollary 5.50 of [Ver10])**.** *Let $w_1, w_2, ..., w_n$ be isotropic random vectors with sub-Gaussian norm upper bounded by κ. Then for every $t > 0$, with probability at least $1 - 2\exp\left\{-c_1t^2\right\}$, the empirical covariance $\widetilde{\mathbf{\Sigma}}$ satisfies,*

$$\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{I}\right\|_2 \leq \max\{\delta,\delta^2\} \quad \textit{where} \quad \delta = c_2\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}$$

where c_1, c_2 are constants depending only on κ.

Remark 2. *For $t = \sqrt{p/c_1}$, we get with probability at least $1 - 2\exp\{-p\}$,*

$$\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{I}\right\|_2 \leq C\sqrt{\frac{p}{n}}$$

where $C = \left\{c_2 + \frac{1}{\sqrt{c_1}}\right\}$, and $n > C^2 p$. Here, C only depends on κ.

Lemma A.1.4 (Corollary 5.52 of [Ver10])**.** *Let $x_1, x_2, ..., x_n$ be random vectors with mean 0 and covariance $\mathbf{\Sigma}$ supported on a centered Euclidean ball of radius $\sqrt{R}$, i.e., $\|x_i\|_2 \leq \sqrt{R}$. For $\epsilon \in (0,1)$ and $c > 0$ an absolute constant, with probability at least $1 - 1/p^2$, for $n > cR\log(p)/(\epsilon^2 \|\mathbf{\Sigma}\|_2)$ the empirical covariance matrix satisfies*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2 \leq \epsilon \|\mathbf{\Sigma}\|_2 .$$

A.2 Additional Experiments

In this section, we provide additional experiments. The overall setting is the same as Section 2.9. The only difference is that we change the sampling distribution of the datasets, which are stated in the title of each plot. As in Section 2.9, SLS estimator outperforms its competitors by a large margin in terms of the computation time.

The results are provided in Figures A.1 and A.2, and Table A.1.

Model	Logistic Regression				Poisson Regression			
Dataset	$\mathbf{\Sigma}\times$Ber($\pm$1)		$\mathbf{\Sigma}\times$Norm(0,1)		$\mathbf{\Sigma}\times\{$Exp(1)-1$\}$		$\mathbf{\Sigma}\times$Norm(0,1)	
Size	$n = 6.0\times10^5$, p=300		$n = 6.0\times10^5$, p=300		$n = 6.0\times10^5$, p=300		$n = 6.0\times10^5$, p=300	
Initialize	Rnd	Ols	Rnd	Ols	Rnd	Ols	Rnd	Ols
Plot	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Method↓	Time in Seconds / Number of Iterations (to reach min test error)							
Sls	6.61/3	2.97/3	9.38/5	4.25/4	14.68/4	2.99/4	6.66/10	4.13/10
Nr	222.21/6	84.08/3	186.33/6	115.76/4	218.1/6	218.9/4	364.63/9	363.4/9
Ns	40.68/10	11.57/3	53.06/9	19.52/4	39.22/6	59.61/4	51.48/10	39.8/10
Bfgs	125.83/33	35.41/9	155.3/48	24.78/8	46.61/20	48.71/12	92.84/36	74.22/38
LBfgs	142.09/38	44.41/12	444.62/143	21.79/7	96.53/39	50.56/12	296.4/111	228.1/117
Gd	409.9/134	79.45/22	1773.1/509	135.62/44	569.1/211	124.31/48	792.3/344	1041.1/366
Agd	177.3/159	43.76/12	359.56/95	53.73/18	157.9/57	63.16/16	74.74/32	62.21/32

Table A.1: Details of the experiments shown in Figures A.1 and A.2.

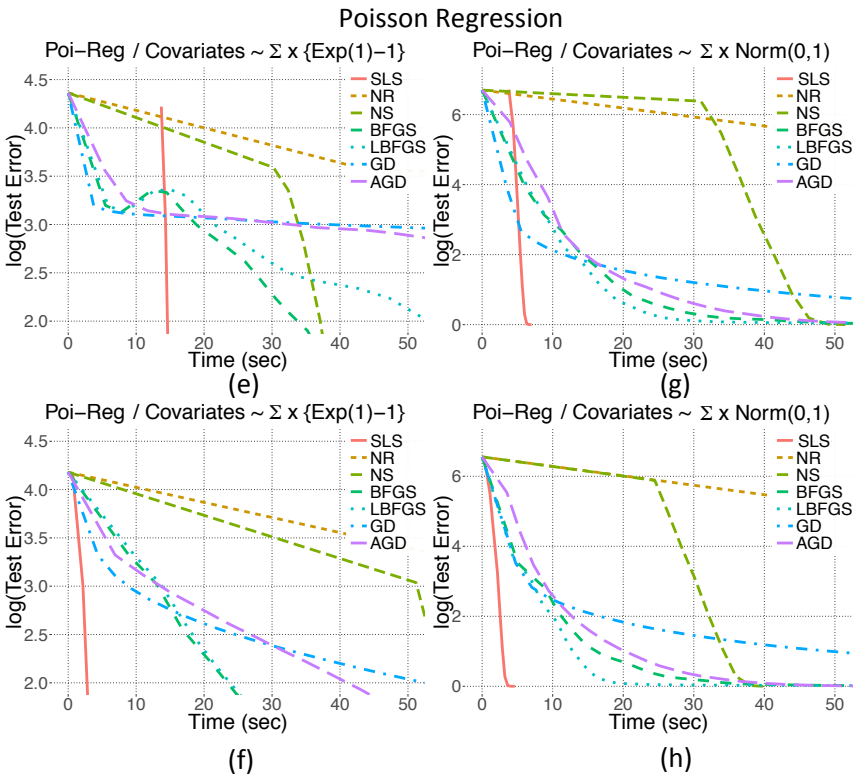

Figure A.1: Additional experiments comparing the performance of SLS to that of MLE obtained with various optimization algorithms on several datasets. SLS is represented with red straight line. The details are provided in Table A.1

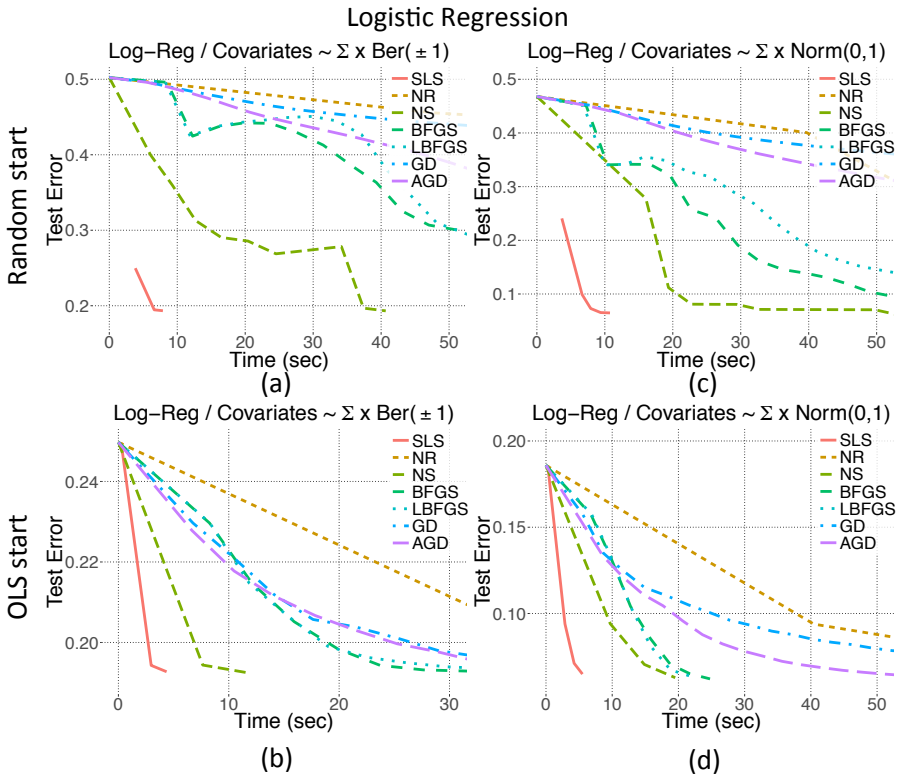

Figure A.2: Additional experiments comparing the performance of SLS to that of MLE obtained with various optimization algorithms on several datasets. SLS is represented with red straight line. The details are provided in Table A.1

Appendix B

Supplement for Chapter 3

B.1 Preliminary Concentration Inequalities

In this section, we provide several concentration bounds that will be useful throughout the proofs.

We state the following Lemmas from [Ver10] for the convenience of the reader (i.e., See Theorem 5.39 and the following remark for sub-Gaussian distributions, and Theorem 5.44 for distributions with arbitrary support):

Lemma B.1.1 ([Ver10])**.** *Let S be an index set and $x_i \in \mathbb{R}^p$ for $i \in S$ be i.i.d. sub-Gaussian random vectors with*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}[x_i x_i^T] = \mathbf{\Sigma}, \qquad \|x_i\|_{\psi_2} \leq K.$$

There exists constants c, C depending only on the sub-Gaussian norm K such that with probability $1 - 2e^{-ct^2}$,

$$\left\| \widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma} \right\|_2 \leq \max\left(\delta, \delta^2\right) \quad \textit{where} \quad \delta = C\sqrt{\frac{p}{|S|}} + \frac{t}{\sqrt{|S|}}.$$

Remark 3. *We are interested in the case where $\delta < 1$, hence the right hand side becomes $\max\left(\delta, \delta^2\right) = \delta$. In most cases, we will simply let $t = \sqrt{p}$ and obtain a bound of order $\sqrt{p/|S|}$ on the right hand side. For this, we need $|S| = \mathcal{O}(C^2 p)$ which is a reasonable assumption in the regime we consider.*

The following lemma is an analogue of Lemma B.1.1 for covariates sampled from arbitrary distributions with bounded support.

Lemma B.1.2 ([Ver10]). *Let S be an index set and $x_i \in \mathbb{R}^p$ for $i \in S$ be i.i.d. random vectors with*

$$\mathbb{E}[x_i] = 0, \qquad \mathbb{E}\left[x_i x_i^T\right] = \mathbf{\Sigma}, \qquad \|x_i\|_2 \leq \sqrt{K} \ \text{a.s.}$$

Then, for some absolute constant c, with probability $1 - pe^{-ct^2}$, we have

$$\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq \max\left(\|\mathbf{\Sigma}\|_2^{1/2}\delta, \delta^2\right) \quad \text{where} \quad \delta = t\sqrt{\frac{K}{|S|}}.$$

Remark 4. *We will choose $t = \sqrt{3\log(p)/c}$ which will provide us with a probability of $1 - 1/p^2$. Therefore, if the sample size is sufficiently large, i.e.,*

$$|S| \geq \frac{3K\log(p)}{c\|\mathbf{\Sigma}\|_2} = \mathcal{O}(K\log(p)/\|\mathbf{\Sigma}\|_2),$$

we can estimate the true covariance matrix quite well for arbitrary distributions with bounded support. In particular, with probability $1 - 1/p^2$, we obtain

$$\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_2 \leq c'\sqrt{\frac{\log(p)}{|S|}},$$

where $c' = \sqrt{3K\|\mathbf{\Sigma}\|_2/c}$.

In the following, we will focus on empirical processes and obtain uniform bounds for proposed Hessian approximation. To that extent, we provide a few basic definitions which will be useful later in the proofs. For a more detailed discussion on the machinery used throughout the next section, we refer reader to [VdV00].

Definition 7. *On a metric space (X, d), for $\epsilon > 0$, $T_\epsilon \subset X$ is called an ϵ-net over X if $\forall x \in X$, $\exists t \in T_\epsilon$ such that $d(x, t) \leq \epsilon$.*

In the following, we will use L_1 distance between two functions f and g, namely $d(f, g) = \int |f - g|$. Note that the same distance definition can be carried to random variables as they are simply real measurable functions. The integral takes the form of expectation.

Definition 8. *Given a function class $\mathcal{F}$, and any two functions l and u (not necessarily in $\mathcal{F}$), the bracket $[l, u]$ is the set of all $f \in \mathcal{F}$ such that $l \leq f \leq u$. A bracket satisfying $l \leq u$ and $\int |u - l| \leq \epsilon$ is called an ϵ-bracket in L_1. The bracketing number $\mathcal{N}(\epsilon, \mathcal{F}, L_1)$ is the minimum number of different ϵ-brackets needed to cover $\mathcal{F}$.*

The preliminary tools presented in this section will be utilized to obtain the concentration results in Section B.2.

B.2 Main Lemmas

B.2.1 Concentration of Covariates With Bounded Support

Lemma B.2.1. *Let $x_i \in \mathbb{R}^p$, for $i = 1, 2, ..., n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, and covariance matrix $\mathbf{\Sigma}$. Further, let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, for sufficiently large n, there exist constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > c_1 \sqrt{\frac{p \log(n)}{n}} \right) \leq c_2 e^{-c_3 p},$$

where the constants depend only on the bound B.

Proof. We start by using the Lipschitz property of the function f, i.e., $\forall \beta, \beta' \in B_p(R)$,

$$\begin{aligned} \left| f(\langle x, \beta \rangle) - f(\langle x, \beta' \rangle) \right| \leq & L \|x\|_2 \|\beta - \beta'\|_2, \\ \leq & L\sqrt{K} \|\beta - \beta'\|_2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwartz. Now let T_Δ be a Δ-net over $B_p(R)$. Then $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that the right hand side of the above inequality is smaller than $\Delta L\sqrt{K}$. Then, we can write

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta' \rangle) - \mathbb{E}[f(\langle x, \beta' \rangle)] \right| + 2\Delta L\sqrt{K}. \quad \text{(B.1)}$$

By choosing

$$\Delta = \frac{\epsilon}{4L\sqrt{K}},$$

and taking supremum over the corresponding β sets on both sides, we obtain the following inequality

$$\sup_{\beta \in B_n(R)} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| \leq \max_{\beta \in \mathcal{T}_\Delta} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| + \frac{\epsilon}{2}.$$

Now, since we have $\|f\|_\infty \leq B$ and for a fixed β and $i = 1, 2, ..., n$, the random variables $f(\langle x_i, \beta \rangle)$ are i.i.d., by the Hoeffding's concentration inequality, we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > \epsilon/2 \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{8B^2} \right).$$

Combining Equation (B.1) with the above result and a union bound, we easily obtain

$$\begin{aligned} &\mathbb{P}\left(\sup_{\beta \in B_n(R)} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > \epsilon \right) \\ &\leq \mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > \epsilon/2 \right) \leq 2|\mathcal{T}_\Delta| \exp\left(-\frac{n\epsilon^2}{8B^2} \right), \end{aligned}$$

where $\Delta = \epsilon / 4L\sqrt{K}$.

Next, we apply Lemma B.4.5 and obtain that

$$|\mathcal{T}_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta} \right)^p = \left(\frac{R\sqrt{p}}{\epsilon / 4L\sqrt{K}} \right)^p.$$

We require that the probability of the desired event is bounded by a quantity that attains an exponential decay with rate $\mathcal{O}(p)$. This can be attained if

$$\epsilon^2 \geq \frac{8B^2 p}{n} \log\left(4eLR\sqrt{K}\sqrt{p}/\epsilon \right).$$

Assuming that n is sufficiently large, and using Lemma B.4.6 with $a = 8B^2 p/n$ and $b = 4eLR\sqrt{Kp}$, we obtain that ϵ should be

$$\epsilon = \sqrt{\frac{4B^2 p}{n} \log\left(\frac{30L^2 R^2 K n}{B^2} \right)} = \mathcal{O}\left(\sqrt{\frac{p \log(n)}{n}} \right).$$

When $n > 30L^2R^2K/B^2$, we obtain

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > 3B\sqrt{\frac{p\log(n)}{n}}\right) \leq 2e^{-p}.$$

□

In the following, we state similar bounds on functions of the following form

$$x \to f(\langle x,\beta\rangle)\langle x,v\rangle^2,$$

which appear in the summation that form the Hessian matrix.

Lemma B.2.2. *Let $x_i \in \mathbb{R}^p$, for $i = 1, ..., n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, and covariance matrix $\mathbf{\Sigma}$. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, for $v \in S^{p-1}$ and sufficiently large n, there exist constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where the constants depend only on the bound B and the radius $\sqrt{K}$.

Proof. As in the proof of Lemma B.2.1, we start by using the Lipschitz property of the function f, i.e., $\forall\beta,\beta' \in B_p(R)$,

$$\begin{aligned}\|f(\langle x,\beta\rangle)\langle x,v\rangle^2 - f(\langle x,\beta'\rangle)\langle x,v\rangle^2\|_2 \leq& L\|x\|_2^3\|\beta-\beta'\|_2,\\ \leq& LK^{1.5}\|\beta-\beta'\|_2.\end{aligned}$$

For a net T_Δ, $\forall\beta \in B_p(R)$, $\exists\beta' \in T_\Delta$ such that right hand side of the above inequality

is smaller than $\Delta LK^{1.5}$. Then, we can write

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta'\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta'\rangle)\langle x,v\rangle^2]\right| \qquad \text{(B.2)} \\
&+ 2\Delta LK^{1.5}. \qquad \text{(B.3)}
\end{aligned}$$

This time, we choose

$$\Delta = \frac{\epsilon}{4LK^{1.5}},$$

and take the supremum over the corresponding feasible β-sets on both sides,

$$\begin{aligned}
&\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| \\
&\leq \max_{\beta\in\mathcal{T}_\Delta}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| + \frac{\epsilon}{2}.
\end{aligned}$$

Now, since we have $\|f\|_\infty \leq B$ and for fixed β and v, $i = 1, 2, ..., n$, $f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2$ are i.i.d. random variables. By the Hoeffding's concentration inequality, we write

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon/2\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8B^2K^2}\right).$$

Using Equation (B.2) and the above result combined with the union bound, we easily obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right) \\
&\leq \mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon/2\right) \\
&\leq 2|\mathcal{T}_\Delta|\exp\left(-\frac{n\epsilon^2}{8B^2K^2}\right),
\end{aligned}$$

where $\Delta = \epsilon/4LK^{1.5}$. Using Lemma B.4.5, we have

$$|\mathcal{T}_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p = \left(\frac{R\sqrt{p}}{\epsilon/4LK^{1.5}}\right)^p.$$

As before, we require that the right hand side of above inequality gets a decay with rate $\mathcal{O}(p)$. Using Lemma B.4.6 with $a = 8B^2K^2p/n$ and $b = 100LRK^{1.5}\sqrt{p}$, we obtain that ϵ should be

$$\epsilon = \sqrt{\frac{4B^2K^2p}{n}\log\left(\frac{50^2L^2R^2Kn}{B^2}\right)} = \mathcal{O}\left(\sqrt{\frac{p\log(n)}{n}}\right).$$

When $n > 50LRK^{1/2}/B$, we obtain

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > 4BK\sqrt{\frac{p\log(n)}{n}}\right) \leq 2e^{-3.2p}.$$

The rate $-3.2p$ will be important later. □

B.2.2 Concentration of Sub-Gaussian Covariates

In this section, we derive the analogues of the Lemmas B.2.1 and B.2.2 for sub-Gaussian covariates. Note that the Lemmas in this section are more general in the sense that they also cover the case where the covariates have bounded support. As a result, the resulting convergence coefficients are worse compared to the previous section.

Lemma B.2.3. *Let $x_i \in \mathbb{R}^p$, for $i = 1, ..., n$, be i.i.d. sub-Gaussian random vectors with mean 0, covariance matrix $\mathbf{\Sigma}$ and sub-Gaussian norm K. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, there exists absolute constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2e^{-c_3p},$$

where the constants depend only on the eigenvalues of $\mathbf{\Sigma}$, bound B and radius R and sub-Gaussian norm K.

Proof. We start by defining the brackets of the form

$$\begin{aligned} l_\beta(x) &= f(\langle x,\beta\rangle) - \epsilon\frac{\|x\|_2}{4\mathbb{E}\left[\|x\|_2\right]}, \\ u_\beta(x) &= f(\langle x,\beta\rangle) + \epsilon\frac{\|x\|_2}{4\mathbb{E}\left[\|x\|_2\right]}. \end{aligned}$$

Observe that the size of bracket $[l_\beta, u_\beta]$ is $\epsilon/2$, i.e., $\mathbb{E}[u_\beta - l_\beta] = \epsilon/2$. Now let T_Δ be a Δ-net over $B_p(R)$ where we use $\Delta = \epsilon/(4L\mathbb{E}\left[\|x\|_2\right])$. Then $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that $f(\langle \cdot, \beta \rangle)$ falls into the bracket $[l_{\beta'}, u_{\beta'}]$. This can be seen by writing out the Lipschitz property of the function f. That is,

$$\begin{aligned} |f(\langle x, \beta \rangle) - f(\langle x, \beta' \rangle)| \leq & L\|x\|_2 \|\beta - \beta'\|_2, \\ \leq & \Delta L \|x\|_2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwartz. Therefore, we conclude that

$$\mathcal{N}(\epsilon/2, \mathcal{F}, L_1) \leq |T_\Delta|$$

for the function class $\mathcal{F} = \{f(\langle \cdot, \beta \rangle) : \beta \in B_p(R)\}$. We further have $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$ such that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \leq \frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)] + \frac{\epsilon}{2}, \\ \frac{1}{n}\sum_{i=1}^n f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \geq \frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) - \mathbb{E}[l_{\beta'}(x)] - \frac{\epsilon}{2}. \end{aligned}$$

Using the above inequalities, we have, $\forall \beta \in B_p(R)$, $\exists \beta' \in T_\Delta$

$$\begin{aligned} &\left\{ \left[\frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)] \right] > \epsilon/2 \right\} \cup \left\{ \left[-\frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) + \mathbb{E}[l_{\beta'}(x)] \right] > \epsilon/2 \right\} \supset \\ &\left\{ \left| \frac{1}{n}\sum_{i=1}^n f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > \epsilon \right\}. \end{aligned}$$

By the union bound, we obtain

$$\begin{aligned} &\mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta} \left[\frac{1}{n}\sum_{i=1}^n u_\beta(x_i) - \mathbb{E}[u_\beta(x)] \right] > \epsilon/2 \right) + \mathbb{P}\left(\max_{\beta \in \mathcal{T}_\Delta} \left[-\frac{1}{n}\sum_{i=1}^n l_\beta(x_i) + \mathbb{E}[l_\beta(x)] \right] > \epsilon/2 \right) \\ &\geq \mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n}\sum_{i=1}^n f(\langle x_i, \beta \rangle) - \mathbb{E}[f(\langle x, \beta \rangle)] \right| > \epsilon \right). \end{aligned} \tag{B.4}$$

In order to complete the proof, we need concentration inequalities for u_β and l_β. We state the following lemma.

Lemma B.2.4. *There exists a constant C depending on the eigenvalues of $\mathbf{\Sigma}$ and B such that, for each $\beta \in B_p(R)$ and for some $0 < \epsilon < 1$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_\beta(x_i) - \mathbb{E}[u_\beta(x)]\right| > \epsilon/2\right) \leq 2e^{-Cn\epsilon^2},$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} l_\beta(x_i) - \mathbb{E}[l_\beta(x)]\right| > \epsilon/2\right) \leq 2e^{-Cn\epsilon^2},$$

where

$$C = \frac{c}{\left(B + \frac{\sqrt{2}K}{4\mu/\sqrt{p}}\right)^2}$$

for an absolute constant c.

Remark 5. *Note that $\mu = \mathbb{E}[\|x\|_2] = \mathcal{O}(\sqrt{p})$ and hence $\mu/\sqrt{p} = \mathcal{O}(1)$.*

Proof. By the relation between sub-Gaussian and sub-exponential norms, we have

$$\begin{aligned} \|\|x\|_2\|_{\psi_2}^2 \leq \|\|x\|_2^2\|_{\psi_1} \leq & \sum_{i=1}^{p} \|x_i^2\|_{\psi_1}, \\ \leq & 2\sum_{i=1}^{p} \|x_i\|_{\psi_2}^2, \\ \leq & 2K^2 p. \end{aligned} \tag{B.5}$$

Therefore $\|x\|_2 - \mathbb{E}[\|x\|_2]$ is a centered sub-Gaussian random variable with sub-Gaussian norm bounded above by $2K\sqrt{2p}$. We have,

$$\mathbb{E}[\|x\|_2] = \mu.$$

Note that μ is actually of order $\sqrt{p}$. Assuming that the left hand side of the above equality is equal to $\sqrt{p}K'$ for some constant $K' > 0$, we can conclude that the random variable

$u_\beta(x) = f(\langle x, \beta\rangle) + \epsilon\frac{\|x\|_2}{4\mathbb{E}[\|x\|_2]}$ is also sub-Gaussian with

$$\begin{aligned}\|u_\beta(x)\|_{\psi_2} \leq & B + \frac{\epsilon}{4\mathbb{E}\left[\|x\|_2\right]}\|\|x\|_2\|_{\psi_2} \\ & \leq B + \frac{\epsilon}{4\sqrt{p}K'}K\sqrt{2p} \\ & \leq B + C'\end{aligned}$$

where $C' = \sqrt{2}K/4K'$ is a constant and we also assumed $\epsilon < 1$. Now, define the function

$$g_\beta(x) = u_\beta(x) - \mathbb{E}[u_\beta(x)].$$

Note that $g_\beta(x)$ is a centered sub-Gaussian random variable with sub-Gaussian norm

$$\|g_\beta(x)\|_{\psi_2} \leq 2B + 2C'.$$

Then, by the Hoeffding-type inequality for the sub-Gaussian random variables, we obtain

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n g_\beta(x_i)\right| > \epsilon/2\right) \leq 2e^{-cn\epsilon^2/(B+C')^2}$$

where c is an absolute constant. The same argument also holds for $l_\beta(x)$. □

Using the above lemma with the union bound over the set T_Δ, we can write

$$\mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i, \beta\rangle) - \mathbb{E}[f(\langle x, \beta\rangle)]\right| > \epsilon\right) \leq 4|T_\Delta|e^{-Cn\epsilon^2}.$$

Since we can also write, by Lemma B.4.5

$$\begin{aligned}|T_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p & \leq \left(\frac{4RL\mathbb{E}[\|x\|_2]\sqrt{p}}{\epsilon}\right)^p, \\ & \leq \left(\frac{4\sqrt{2}RLKp}{\epsilon}\right)^p,\end{aligned}$$

and we observe that, for the constant $c' = 4\sqrt{2}RLK$,

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > \epsilon\right) \leq 4\left(\frac{4\sqrt{2}RLKp}{\epsilon}\right)^p e^{-Cn\epsilon^2},$$
$$= 4\exp\left\{p\log(c'p/\epsilon) - Cn\epsilon^2\right\}.$$

We will obtain an exponential decay of order p on the right hand side. For some constant h depending on n and p, if we choose $\epsilon = hp$, we need

$$h^2 \geq \frac{1}{Cnp}\log(c'/h).$$

By the Lemma B.4.6, choosing $h^2 = \log(2c'^2Cnp)/(2Cnp)$, we satisfy the above requirement. Note that for n large enough, the condition of the lemma is easily satisfied. Hence, for

$$\epsilon^2 = \frac{p\log(2c'^2Cnp)}{2Cn} = \mathcal{O}\left(\frac{p\log(n)}{n}\right),$$

we obtain that there exists constants c_1, c_2, c_3 such that

$$\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle) - \mathbb{E}[f(\langle x,\beta\rangle)]\right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where

$$\begin{aligned} c_1 =& \frac{3\left(B + \frac{\sqrt{2}K}{4\sqrt{\mathrm{Tr}(\mathbf{\Sigma})/p - 16K^2}}\right)^2}{2c}, \\ c_2 =& 4, \\ c_3 =& \frac{1}{2}\log(7) \leq \frac{1}{2}\log(\log(64R^2L^2K^2C) + 6\log(p)). \end{aligned}$$

when $p > e$ and $64R^2L^2K^2C > e$.

□

In the following, we state the concentration results on the unbounded functions of the form

$$x \to f(\langle x,\beta\rangle)\langle x,v\rangle^2.$$

Functions of this type form the summands of the Hessian matrix in GLMs.

Lemma B.2.5. *Let x_i, for $i = 1, ..., n$, be i.i.d sub-Gaussian random variables with mean 0, covariance matrix $\mathbf{\Sigma}$ and sub-Gaussian norm K. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Further, let $v \in \mathbb{R}^p$ such that $\|v\|_2 = 1$. Then, for n, p sufficiently large satisfying*

$$n^{0.2}/\log(n) \gtrsim p,$$

there exist constants c_1, c_2 depending on L, B, R and the eigenvalues of $\mathbf{\Sigma}$ such that, we have

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \beta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \beta\rangle)\langle x, v\rangle^2] \right| > c_1 \sqrt{\frac{p}{n^{0.2}} \log(n)} \right) \leq c_2 e^{-p}.$$

Proof. We define the brackets of the form

$$\begin{aligned} l_\beta(x) =& f(\langle x, \beta\rangle)\langle x, v\rangle^2 - \epsilon \frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]}, \\ u_\beta(x) =& f(\langle x, \beta\rangle)\langle x, v\rangle^2 + \epsilon \frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]}, \end{aligned} \tag{B.6}$$

and we observe that the bracket $[l_\beta, u_\beta]$ has size $\epsilon/2$ in L_1, that is,

$$\mathbb{E}\left[|u_\beta(x) - l_\beta(x)|\right] = \epsilon/2.$$

Next, for the following constant

$$\Delta = \frac{\epsilon}{4L\mathbb{E}\left[\|x\|_2^3\right]},$$

we define a Δ-net over $B_p(R)$ and call it $\mathcal{T}_\Delta$. Then, $\forall \beta \in B_p(R)$, $\exists \beta' \in \mathcal{T}_\Delta$ such that $f(\langle \cdot, \beta\rangle)\langle \cdot, v\rangle^2$ belongs to the bracket $[l_{\beta'}, u_{\beta'}]$. This can be seen by writing the Lipschitz

continuity of the function f, i.e.,

$$\begin{aligned}\left|f(\langle x,\beta\rangle)\langle x,v\rangle^2 - f(\langle x,\beta'\rangle)\langle x,v\rangle^2\right| =& \langle x,v\rangle^2 \left|\left\{f(\langle x,\beta\rangle) - f(\langle x,\beta'\rangle)\right\}\right|, \\ \leq& L\|x\|_2^2 \ \|v\|_2^2 \ \left|\langle x,\beta-\beta'\rangle\right|, \\ \leq& L\|x\|_2^3\|\beta-\beta'\|_2, \\ \leq& \Delta L\|x\|_2^3,\end{aligned}$$

where we used Cauchy-Schwartz to obtain the above inequalities. Hence, we may conclude that for the bracketing functions given in Equation (B.6), the corresponding bracketing number of the function class

$$\mathcal{F} = \{f(\langle\cdot,\beta\rangle)\langle\cdot,v\rangle^2 : \beta \in B_p(R)\}$$

is bounded above by the covering number of the ball of radius R for the given scale $\Delta = \epsilon/(4L\mathbb{E}\left[\|x\|_2^3\right])$, i.e.,

$$\mathcal{N}(\epsilon/2, \mathcal{F}, L_1) \leq |\mathcal{T}_\Delta|.$$

Next, we will upper bound the target probability using the bracketing functions u_β, l_β. We have $\forall\beta \in B_p(R)$, $\exists\beta' \in \mathcal{T}_\Delta$ such that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2] \leq \frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)] + \frac{\epsilon}{2}, \\ \frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2] \geq \frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) - \mathbb{E}[l_{\beta'}(x)] - \frac{\epsilon}{2}.\end{aligned}$$

Using the above inequalities, $\forall\beta \in B_p(R)$, $\exists\beta' \in \mathcal{T}_\Delta$, we can write

$$\begin{aligned}&\left\{\left[\frac{1}{n}\sum_{i=1}^n u_{\beta'}(x_i) - \mathbb{E}[u_{\beta'}(x)]\right] > \epsilon/2\right\} \cup \left\{\left[-\frac{1}{n}\sum_{i=1}^n l_{\beta'}(x_i) + \mathbb{E}[l_{\beta'}(x)]\right] > \epsilon/2\right\} \supset \\ &\left\{\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right\}.\end{aligned}$$

Hence, by the union bound, we obtain

$$\mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left[\frac{1}{n}\sum_{i=1}^n u_\beta(x_i)-\mathbb{E}[u_\beta(x)]\right]>\epsilon/2\right)+\mathbb{P}\left(\max_{\beta\in\mathcal{T}_\Delta}\left[-\frac{1}{n}\sum_{i=1}^n l_\beta(x_i)+\mathbb{E}[l_\beta(x)]\right]>\epsilon/2\right)$$
$$\geq \mathbb{P}\left(\sup_{\beta\in B_p(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2-\mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right|>\epsilon\right). \tag{B.7}$$

In order to complete the proof, we need one-sided concentration inequalities for u_β and l_β. Handling these functions is somewhat tedious since $\|x\|_2^3$ terms do not concentrate nicely. We state the following lemma.

Lemma B.2.6. *For given $\alpha,\epsilon>0$, and n sufficiently large such that,*

$$\nu(n^\alpha,p,\epsilon,B,K,\mathbf{\Sigma})<\epsilon/4$$

where

$$\nu(n^\alpha,p,\epsilon,B,K,\mathbf{\Sigma}) =: 2\left(n^\alpha+\frac{6BK^2p}{c}\right)\exp\left(-c\frac{n^\alpha}{6BK^2p}\right)+2\left\{n^\alpha+\frac{3K^2p}{c\mathrm{Tr}\left(\mathbf{\Sigma}\right)}n^{\alpha/3}\epsilon^{2/3}\right.$$
$$\left.+\frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\right\}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).$$

Then, there exists constants c',c'',c''' depending on the eigenvalues of $\mathbf{\Sigma}$, B and K such that $\forall\beta$, we have,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n u_\beta(x_i)-\mathbb{E}[u_\beta(x)]>\epsilon/2\right)\leq 2\exp\left(-c'n^\alpha/p\right)$$
$$+2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right)+\exp\left(-c'''n^{1-2\alpha}\epsilon^2\right),$$

and

$$\mathbb{P}\left(-\frac{1}{n}\sum_{i=1}^n l_\beta(x_i)+\mathbb{E}[l_\beta(x)]>\epsilon/2\right)\leq 2\exp\left(-c'n^\alpha/p\right)+$$
$$2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right)+\exp\left(-c'''n^{1-2\alpha}\epsilon^2\right).$$

Proof. For the sake of simplicity, we define the functions

$$\tilde{u}_\beta(w) = u_\beta(w) - \mathbb{E}[u_\beta(x)],$$
$$\tilde{l}_\beta(w) = l_\beta(w) - \mathbb{E}[l_\beta(x)].$$

We will derive the result for the upper bracket, $\tilde{u}$, and skip the proof for the lower bracket $\tilde{l}$ as it follows from the same steps. We write,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \epsilon/2\right) \leq & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \epsilon/2,\ \max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| < n^\alpha\right) \\ & + \mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| \geq n^\alpha\right). \end{aligned} \tag{B.8}$$

We need to bound the right hand side of the above equation. For the second term, since $\tilde{u}_\beta(x_i)$'s are i.i.d. centered random variables, we have

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| \geq n^\alpha\right) =& 1 - \mathbb{P}\left(\max_{1\leq i\leq n}|\tilde{u}_\beta(x_i)| < n^\alpha\right), \\ =& 1 - \mathbb{P}\left(|\tilde{u}_\beta(x)| < n^\alpha\right)^n, \\ =& 1 - \left(1 - \mathbb{P}\left(|\tilde{u}_\beta(x)| \geq n^\alpha\right)\right)^n, \\ \leq& n\mathbb{P}\left(|\tilde{u}_\beta(x)| \geq n^\alpha\right).\end{aligned}$$

Also, note that

$$\begin{aligned}|\tilde{u}_\beta(x)| \leq& B\|x\|_2^2 + \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} + \mathbb{E}[u_\beta(x)], \\ \leq& B\|x\|_2^2 + \epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} + B\lambda_{\max}(\mathbf{\Sigma}) + \epsilon/4.\end{aligned}$$

Therefore, if $t > 3B\lambda_{\max}(\mathbf{\Sigma})$ and for ϵ small, we can write

$$\{|\tilde{u}_\beta(x)| > t\} \subset \left\{B\|x\|_2^2 > t/3\right\} \cup \left\{\epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} > t/3\right\}. \tag{B.9}$$

Since x is a sub-Gaussian random variable with $\|x\|_{\psi_2} = K$, we have

$$K = \sup_{w \in S^{p-1}} \|\langle w, x \rangle\|_{\psi_2} = \|x\|_{\psi_2}.$$

Using this and the relation between sub-Gaussian and sub-exponential norms as in Equation (B.5), we have $\|\|x\|_2\|_{\psi_2}^2 \leq 2K^2 p$. This provides the following tail bound for $\|x\|_2$,

$$\mathbb{P}\left(\|x\|_2 > s\right) \leq 2\exp\left(-\frac{cs^2}{2pK^2}\right), \tag{B.10}$$

where c is an absolute constant. Using the above tail bound, we can write,

$$\mathbb{P}\left(\|x\|_2^2 > \frac{1}{3B}t\right) \leq 2\exp\left(-c\frac{t}{6BK^2p}\right).$$

For the next term in Equation (B.9), we need a lower bound for $\mathbb{E}\left[\|x\|_2^3\right]$. We use a modified version of the Hölder's inequality and obtain

$$\mathbb{E}\left[\|x\|_2^3\right] \geq \mathbb{E}\left[\|x\|_2^2\right]^{3/2} = \mathrm{Tr}(\mathbf{\Sigma})^{3/2}.$$

Using the above inequality, we can write

$$\begin{aligned}
\mathbb{P}\left(\epsilon\frac{\|x\|_2^3}{4\mathbb{E}\left[\|x\|_2^3\right]} > t/3\right) \leq & \mathbb{P}\left(\|x\|_2^3 > \frac{4}{3\epsilon}\mathrm{Tr}(\mathbf{\Sigma})^{3/2}t\right), \\
= & \mathbb{P}\left(\|x\|_2 > \left(\frac{4t}{3\epsilon}\right)^{1/3}\mathrm{Tr}(\mathbf{\Sigma})^{1/2}\right), \\
\leq & 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right),
\end{aligned}$$

where c is the same absolute constant as in Equation (B.10).

Now for $\alpha > 0$ such that $t = n^\alpha > 3B\lambda_{\max}(\mathbf{\Sigma})$ (we will justify this assumption for a particular choice of α later), we combine the above results,

$$\mathbb{P}\left(|\tilde{u}_\beta(x)| > t\right) \leq 2\exp\left(-c\frac{t}{6BK^2p}\right) + 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right). \tag{B.11}$$

Next, we focus on the first term in Equation (B.8). Let $\mu = \mathbb{E}[\tilde{u}_\beta(x)\mathbb{1}_{\{|\tilde{u}_\beta(x)| < n^\alpha\}}]$, and

write

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \frac{\epsilon}{2};\ \max_{1\le i\le n}|\tilde{u}_\beta(x_i)| < n^\alpha\right)\\
&\le \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i)\mathbb{1}_{\{|\tilde{u}_\beta(x_i)|<n^\alpha\}} > \frac{\epsilon}{2}\right),\\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i)\mathbb{1}_{\{|\tilde{u}_\beta(x_i)|<n^\alpha\}} - \mu > \frac{\epsilon}{2} - \mu\right)\\
&\le \exp\left\{-\frac{n^{1-2\alpha}}{2}\left(\frac{\epsilon}{2}-\mu\right)^2\right\},
\end{aligned}$$

where we used the Hoeffding's concentration inequality for the bounded random variables. Further, note that

$$0 = \mathbb{E}[\tilde{u}_\beta(x)] = \mu + \mathbb{E}\left[\tilde{u}_\beta(x)\mathbb{1}_{\{|\tilde{u}_\beta(x)|>n^\alpha\}}\right].$$

By Lemma B.4.2, we can write

$$|\mu| = \left|\mathbb{E}\left[\tilde{u}_\beta(x)\mathbb{1}_{\{|\tilde{u}_\beta(x)|>n^\alpha\}}\right]\right| \le n^\alpha\mathbb{P}(|\tilde{u}_\beta(x)|>n^\alpha) + \int_{n^\alpha}^{\infty}\mathbb{P}(|\tilde{u}_\beta(x)|>t)dt.$$

The first term on the right hand side can be easily bounded by using Equation (B.11), i.e.,

$$n^\alpha\mathbb{P}(|\tilde{u}_\beta(x)|>n^\alpha) \le 2n^\alpha\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) + 2n^\alpha\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).$$

For the second term, using Equation (B.11) once again, we obtain

$$\begin{aligned}
&\int_{n^\alpha}^{\infty}\mathbb{P}(|\tilde{u}_\beta(x)|>t)dt\\
\le& 2\int_{n^\alpha}^{\infty}\exp\left(-c\frac{t}{6BK^2p}\right)dt + 2\int_{n^\alpha}^{\infty}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right)dt,\\
=& \frac{12BK^2p}{c}\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) + 2\int_{n^\alpha}^{\infty}\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right)dt.
\end{aligned}$$

Next, we apply Lemma B.4.3 to bound the second term on the right hand side. That is, we

have

$$\begin{aligned}
&\int_{n^\alpha}^{\infty} \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(t/\epsilon)^{2/3}}{2K^2p}\right) dt \\
&\leq \left\{\frac{3K^2p}{c\mathrm{Tr}\left(\mathbf{\Sigma}\right)} n^{\alpha/3}\epsilon^{2/3} + \frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\right\} \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right).
\end{aligned}$$

Combining the above results, we can write

$$\begin{aligned}
|\mu| \leq& 2\left(n^\alpha + \frac{6BK^2p}{c}\right)\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) \\
&+ 2\left\{n^\alpha + \frac{3K^2p}{c\mathrm{Tr}\left(\mathbf{\Sigma}\right)} n^{\alpha/3}\epsilon^{2/3} + \frac{3K^4p^2}{c^2\mathrm{Tr}(\mathbf{\Sigma})^2}\epsilon^{4/3}n^{-\alpha/3}\right\} \exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right), \\
=:& \nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma}).
\end{aligned}$$

Notice that, the upper bound on $|\mu|$, namely $\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma})$, is close to 0 when n is large. This is because of exponentially decaying functions that dominates the other terms. We assume that n is sufficiently large that the upper bound for $|\mu|$ is less than $\epsilon/4$. For the value of α, we will choose $\alpha = 0.4$ later in the proof.

Applying this bounds in Equation (B.8), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{u}_\beta(x_i) > \epsilon/2\right) \\
&\leq 2\exp\left(-c\frac{n^\alpha}{6BK^2p}\right) \\
&+ 2\exp\left(-c\frac{\mathrm{Tr}(\mathbf{\Sigma})(n^\alpha/\epsilon)^{2/3}}{2K^2p}\right) + \exp\left(-\frac{n^{1-2\alpha}}{32}\epsilon^2\right), \\
&= 2\exp\left(-c'n^\alpha/p\right) + 2\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right) + \exp\left(-c'''n^{1-2\alpha}\epsilon^2\right),
\end{aligned}$$

where

$$\begin{aligned}
c' =& \frac{c}{6BK^2}, \\
c'' =& \frac{c\mathrm{Tr}(\mathbf{\Sigma})/p}{2K^2} \geq \frac{c\lambda_{\min}(\mathbf{\Sigma})}{2K^2}, \\
c''' =& \frac{1}{32}.
\end{aligned}$$

Hence, the proof is completed for the upper bracket.

The proof for the lower brackets $l_\beta(x)$ follows from exactly the same steps and omitted here. □

Applying the above lemma on Equation (B.7), for $\alpha > 0$, we obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{\beta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^{n} f(\langle x_i,\beta\rangle)\langle x_i,v\rangle^2 - \mathbb{E}[f(\langle x,\beta\rangle)\langle x,v\rangle^2]\right| > \epsilon\right) \\ &\leq 4|T_\Delta|\exp\left(-c'n^\alpha/p\right) + 4|T_\Delta|\exp\left(-c''n^{2\alpha/3}\epsilon^{-2/3}\right) + 2|T_\Delta|\exp\left(-c'''n^{1-2\alpha}\epsilon^2\right).\end{aligned} \tag{B.12}$$

Observe that we can write, by Lemma B.4.5

$$|T_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p = \left(\frac{4\sqrt{p}RL\mathbb{E}[\|x\|_2^3]}{\epsilon}\right)^p.$$

Also, recall that $\|x\|_2$ was a sub-Gaussian random variable with $\|\|x\|_2\|_{\psi_2} \leq K\sqrt{2p}$. Using the definition of sub-Gaussian norm, we have

$$\begin{aligned}\frac{1}{\sqrt{3}}\,\mathbb{E}[\|x\|_2^3]^{1/3} \leq &\|\|x\|_2\|_{\psi_2} \leq \sqrt{2p}\,K, \\ &\implies \mathbb{E}[\|x\|_2^3] \leq 15K^3p^{3/2}.\end{aligned}$$

Therefore, we have $\mathbb{E}[\|x\|_2^3] = \mathcal{O}(p^{3/2})$ (recall that we had a lower bound of the same order). We define a constant K', and as ϵ is small, we have

$$|T_\Delta| \leq \left(\frac{60RLK^3p^2}{\epsilon}\right)^p = \left(\frac{K'p^2}{\epsilon}\right)^p,$$

where we let $K' = 60RLK^3$. We will show that each term on the right hand side of Equation (B.12) decays exponentially with a rate of order p. For the first term, for $s > 0$, we write

$$\begin{aligned}|T_\Delta|\exp\left(-c'n^\alpha/p\right) &= \exp\left(-c'n^\alpha/p + p\log(K') + 2p\log(p) + p\log(\epsilon^{-1})\right), \\ &\leq \exp\left(-c'n^\alpha/p + 2p\log(K'p/\epsilon)\right).\end{aligned} \tag{B.13}$$

Similarly for the second and third terms, we write

$$\begin{aligned} |T_\Delta| \exp\left(-c'' n^{2\alpha/3}\epsilon^{-2/3}\right) &\leq \exp\left(-c'' n^{2\alpha/3}\epsilon^{-2/3} + 2p\log(K'p/\epsilon)\right), \\ |T_\Delta| \exp\left(-c''' n^{1-2\alpha}\epsilon^{2}\right) &\leq \exp\left(-c''' n^{1-2\alpha}\epsilon^{2} + 2p\log(K'p/\epsilon)\right). \end{aligned} \tag{B.14}$$

We will seek values for ϵ and α to obtain an exponential decay with rate p on the right sides of Equations B.13 and B.14. That is, we need

$$\begin{aligned} c' n^{\alpha}/p &\geq 2p\log(K''p/\epsilon), \\ c'' n^{2\alpha/3} &\geq 2p\log(K''p/\epsilon)\epsilon^{2/3}, \\ c''' n^{1-2\alpha}\epsilon^{2} &\geq 2p\log(K''p/\epsilon), \end{aligned} \tag{B.15}$$

where $K'' = eK'$.

We apply Lemma B.4.6 for the last inequality in Equation (B.15). That is,

$$\begin{aligned} \epsilon^2 &= \frac{p}{c''' n^{1-2\alpha}} \log\left(c''' K''^2 p n^{1-2\alpha}\right), \\ &= \mathcal{O}\left(\frac{p}{n^{1-2\alpha}}\log(n)\right). \end{aligned} \tag{B.16}$$

where we assume that n is sufficiently large. The above statement holds for $\alpha < 1/2$.

In the following, we choose $\alpha = 0.4$ and use the assumption that

$$n^{0.2}/\log(n) \gtrsim p, \tag{B.17}$$

which provides $\epsilon < 1$. Note that this choice of α also justifies the assumption used to derive Equation (B.11). One can easily check that $\alpha = 0.4$ implies that the first and the second statements in Equation (B.15) are satisfied for sufficiently large n.

It remains to check whether $\nu(n^\alpha, p, \epsilon, B, K, \mathbf{\Sigma}) < \epsilon/4$ (in Lemma B.2.6) for this particular choice of α and ϵ. It suffices to consider only the dominant terms in the definition of

ν. We use the assumption on n, p and write

$$\begin{aligned}\nu(n^{0.4}, p, \epsilon, B, K, \mathbf{\Sigma}) \lesssim & n^{0.4} \exp\left(-\frac{cn^{0.4}}{6BK^2p}\right) + n^{0.4} \exp\left(-\frac{c\mathrm{Tr}(\mathbf{\Sigma})/p\ n^{0.8/3}}{2K^2}\right), \\ \lesssim & n^{0.4} \exp\left(-\frac{c}{6BK^2} n^{0.2}\right) + n^{0.4} \exp\left(-\frac{c\lambda_{\min}(\mathbf{\Sigma})}{2K^2} n^{0.8/3}\right).\end{aligned} \tag{B.18}$$

For n sufficiently large, due to exponential decay in $n^{0.2}$, the above quantity can be made arbitrarily small. Hence, for some constants c_1, c_2, we obtain

$$\mathbb{P}\left(\sup_{\beta \in B_p(R)} \left| \frac{1}{n} \sum_{i=1}^{n} f(\langle x_i, \beta \rangle) \langle x_i, v \rangle^2 - \mathbb{E}[f(\langle x, \beta \rangle) \langle x, v \rangle^2] \right| > c_1 \sqrt{\frac{p}{n^{0.2}} \log(n)}\right) \leq c_2 e^{-p}.$$

□

B.3 Local Step Size Selection

This section provides a heuristic calculation for choosing a local step size when eigenvalue thresholding is applied to the Newton-Stein method. We carry our analysis from Equation (3.14). The optimal local step size would be

$$\gamma_* = \underset{\gamma}{\mathrm{argmin}} \left\| I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi \right\|_2. \tag{B.19}$$

Defining the following matrix,

$$\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t) = \int_0^1 \boldsymbol{\nabla}^2_\beta l(\beta_* + \xi(\hat{\beta}^t - \beta_*)) d\xi, \tag{B.20}$$

and we write the governing term as

$$\left\| I - \gamma \mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t) \right\|_2. \tag{B.21}$$

The above function is piecewise linear in γ and it can be minimized by setting

$$\gamma_* = \frac{2}{\lambda_1\left(\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)\right) + \lambda_p\left(\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)\right)}. \tag{B.22}$$

Since we don't have access to the optimal value β_*, we cannot determine the exact value of $\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$. Hence, we will assume that $\boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$ and the current estimate are close.

In the regime $n \gg p$, and by our construction of the scaling matrix $\mathbf{Q}^t$, we have

$$\mathbf{Q}^t \approx \left[\mathbb{E}[xx^T \Psi^{(2)}(\langle x, \hat{\beta}^t \rangle)]\right]^{-1} \quad \text{and}$$
$$\boldsymbol{\nabla}^2_\beta \ell(\hat{\beta}^t) \approx \mathbb{E}[xx^T \Psi^{(2)}(\langle x, \hat{\beta}^t \rangle)].$$

The crucial observation is that the eigenvalue thresholding suggested in [EM15] estimates the smallest eigenvalue with $(r+1)$-th eigenvalue (say $\hat{\sigma}^2$) which overestimates true value (say σ^2) in general. Even though the largest eigenvalue of $\mathbf{Q}^t \boldsymbol{\nabla}^2_\beta \tilde{\ell}(\hat{\beta}^t)$ will be close to 1, the smallest value will be $\sigma^2/\hat{\sigma}^2$. This will make the optimal step size larger than 1. Hence, we suggest

$$\gamma = \frac{2}{1 + \sigma^2/\hat{\sigma}^2}, \tag{B.23}$$

if σ^2 were known. We also have, by the Weyl's inequality,

$$\left|\hat{\sigma}^2 - \sigma^2\right| \leq \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \leq C\sqrt{\frac{p}{|S|}}, \tag{B.24}$$

with high probability. Whenever r is less than $p/2$, we suggest to use

$$\gamma = \frac{2}{1 + \frac{\hat{\sigma}^2 - \mathcal{O}(\sqrt{p/|S|})}{\hat{\sigma}^2}}, \tag{B.25}$$

if σ^2 is unknown.

B.4 Useful Lemmas

Lemma B.4.1. *Let* Γ *denote the* Gamma *function. Then, for* $r \in (0,1)$*, we have*

$$z^{1-r} < \frac{\Gamma(z+1)}{\Gamma(z+r)} < (1+z)^{1-r}.$$

Lemma B.4.2. *Let* Z *be a random variable with a density function* f *and cumulative*

distribution function F. If $F^C = 1 - F$, then,

$$\left|\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}]\right| \leq t\mathbb{P}(|Z| > t) + \int_t^\infty \mathbb{P}(|Z| > z)dz.$$

Proof. We write,

$$\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}] = \int_t^\infty zf(z)dz + \int_{-\infty}^{-t} zf(z)dz. \tag{B.26}$$

Using integration by parts, we obtain

$$\begin{aligned}\int zf(z)dz =& - zF^C(z) + \int F^C(z)dz,\\ =& zF(z) - \int F(z)dz.\end{aligned}$$

Since $\lim_{z\to\infty} zF^C(z) = \lim_{z\to-\infty} zF(z) = 0$, we have

$$\begin{aligned}\int_t^\infty zf(z)dz =& tF^C(t) + \int_t^\infty F^C(z)dz,\\ \int_{-\infty}^{-t} zf(z)dz =& - tF(-t) - \int_{-\infty}^{-t} F(z)dz,\\ =& - tF(-t) - \int_t^\infty F(-z)dz.\end{aligned}$$

Hence, we obtain the following bound,

$$\begin{aligned}\left|\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}]\right| =& \left|tF^C(t) + \int_t^\infty F^C(z)dz - tF(-t) - \int_t^\infty F(-z)dz\right|,\\ \leq& t\left(F^C(t) + F(-t)\right) + \left(\int_t^\infty F^C(z) + F(-z)dz\right),\\ \leq& t\mathbb{P}(|Z| > t) + \int_t^\infty \mathbb{P}(|Z| > z)dz.\end{aligned}$$

□

Lemma B.4.3. *For positive constants c_1, c_2, we have*

$$\int_{c_1}^\infty e^{-c_2t^{2/3}}dt \leq \left\{\frac{3c_1^{1/3}}{2c_2} + \frac{3}{4c_2^2c_1^{1/3}}\right\} e^{-c_2c_1^{2/3}} \tag{B.27}$$

Proof. By the change of variables $t^{2/3} = x^2$, we get

$$\int_{c_1}^{\infty} e^{-c_2 t^{2/3}} dt = 3 \int_{c_1^{1/3}}^{\infty} x^2 e^{-c_2 x^2} dx. \tag{B.28}$$

Next, we notice that

$$de^{-c_2 x^2} = -2c_2 x e^{-c_2 x^2} dx. \tag{B.29}$$

Hence, using the integration by parts, we have

$$\int_{c_1}^{\infty} e^{-c_2 t^{2/3}} dt = \frac{3}{2c_2} \left\{ c_1^{1/3} e^{-c_2 c_1^{2/3}} + \int_{c_1^{1/3}}^{\infty} e^{-c_2 x^2} dx \right\}. \tag{B.30}$$

We will find an upper bound on the second term. Using the change of variables, $x = y + c_1^{1/3}$, we obtain

$$\begin{aligned} \int_{c_1^{1/3}}^{\infty} e^{-c_2 x^2} dx &= \int_0^{\infty} e^{-c_2 \left(y + c_1^{1/3}\right)^2} dy, \\ &\leq e^{-c_2 c_1^{2/3}} \int_0^{\infty} e^{-2c_2 y c_1^{1/3}} dy, \\ &= \frac{e^{-c_2 c_1^{2/3}}}{2c_2 c_1^{1/3}}. \end{aligned}$$

Combining the above results, we complete the proof. □

Lemma B.4.4 ([Ver10])**.** *Let X be a symmetric $p \times p$ matrix, and let T_ϵ be an ϵ-net over S^{p-1}. Then,*

$$\|X\|_2 \leq \frac{1}{1 - 2\epsilon} \sup_{v \in T_\epsilon} |\langle Xv, v \rangle| . \tag{B.31}$$

Lemma B.4.5. *Let $B_p(R) \subset \mathbb{R}^p$ be the ball of radius R centered at the origin and T_ϵ be an ϵ-net over $B_p(R)$. Then,*

$$|T_\epsilon| \leq \left(\frac{R\sqrt{p}}{\epsilon} \right)^p . \tag{B.32}$$

Proof. A similar proof appears in [VdV00]. The set $B_p(R)$ can be contained in a p-dimensional cube of size $2R$. Consider a grid over this cube with mesh width $2\epsilon/\sqrt{p}$. Then $B_p(R)$ can be covered with at most $(2R/(2\epsilon/\sqrt{p}))^p$ many cubes of edge length $2\epsilon/\sqrt{p}$. If ones takes the projection of the centers of such cubes onto $B_p(R)$ and considers the circumscribed balls of radius ϵ, we may conclude that $B_p(R)$ can be covered with at most

$$\left(\frac{2R}{2\epsilon/\sqrt{p}}\right)^p$$

many balls of radius ϵ. □

Lemma B.4.6. *For $a, b > 0$, and ϵ satisfying*

$$\epsilon = \left\{\frac{a}{2}\log\left(\frac{2b^2}{a}\right)\right\}^{1/2} \qquad \text{and} \qquad \frac{2}{a}b^2 > e, \tag{B.33}$$

we have $\epsilon^2 \geq a\log(b/\epsilon)$. Moreover, the gap in the inequality can be written as

$$\epsilon^2 - a\log(b/\epsilon) = \frac{a}{2}\log\log\left(\frac{2b^2}{a}\right). \tag{B.34}$$

Proof. Since $a, b > 0$ and $x \to e^x$ is a monotone increasing function, the above inequality condition is equivalent to

$$\frac{2\epsilon^2}{a}e^{\frac{2\epsilon^2}{a}} \geq \frac{2b^2}{a}. \tag{B.35}$$

Now, we use the function $f(w) = we^w$ for $w > 0$ (in fact this function is well-known by the name Lambert W function). f is continuous and invertible on $[0, \infty)$. Note that f^{-1} is also a continuous and increasing function for $w > 0$. Therefore, we have

$$\epsilon^2 \geq \frac{a}{2}f^{-1}\left(\frac{2b^2}{a}\right) \tag{B.36}$$

Observe that the smallest possible value for ϵ would be simply the square root of $af^{-1}\left(2b^2/a\right)/2$. For simplicity, we will obtain a more interpretable expression for ϵ. By the definition of f^{-1}, we have

$$\log(f^{-1}(y)) + f^{-1}(y) = \log(y). \tag{B.37}$$

Since the condition on a and b enforces $f^{-1}(y)$ to be larger than 1, we obtain the following simple inequality that

$$f^{-1}(y) \leq \log(y). \tag{B.38}$$

Using the above inequality, if ϵ satisfies

$$\epsilon^2 = \frac{a}{2} \log\left(\frac{2b^2}{a}\right), \tag{B.39}$$

we obtain the desired result,

$$\epsilon^2 \geq a \log(b/\epsilon). \tag{B.40}$$

□

Appendix C

Supplement for Chapter 4

C.1 Properties of composite convergence

In the previous sections, we showed that NewSamp gets a composite convergence rate, i.e., the ℓ_2 distance from the current iterate to the optimal value can be bounded by the sum of a linearly and a quadratically converging term. We study such convergence rates assuming the coefficients do not change at each iteration t. Denote by Δ_t, the aforementioned ℓ_2 distance at iteration step t, i.e.,

$$\Delta_t = \|\hat{\beta}^t - \theta_*\|_2, \tag{C.1}$$

and assume that the algorithm gets a composite convergence rate as

$$\forall t \geq 0, \qquad \Delta_{t+1} \leq \xi_1 \Delta_t + \xi_2 \Delta_t^2, \tag{C.2}$$

where $\xi_1, \xi_2 > 0$ denote the coefficients of linearly and quadratically converging terms, respectively.

C.1.1 Local asymptotic rate

We state the following theorem on the local convergence properties of compositely converging algorithms.

Lemma C.1.1. *For a compositely converging algorithm as in Equation* (C.1) *with coefficients* $1 > \xi_1, \xi_2 > 0$*, if the initial distance* Δ_0 *satisfies* $\Delta_0 < (1 - \xi_1)/\xi_2$*, then we have*

$$\limsup_{t \to \infty} -\frac{1}{t} \log(\Delta_t) \leq -\log(\xi_1). \tag{C.3}$$

The above theorem states that the local convergence of a compositely converging algorithm will be dominated by the linear term.

Proof of Lemma C.1.1. The condition on the initial point implies that $\Delta_t \to 0$ as $t \to \infty$. Hence, for any given $\delta > 0$, there exists a positive integer T such that $\forall t \geq T$, we have $\Delta_t < \delta/\xi_2$. For such values of t, we write

$$\xi_1 + \xi_2 \Delta_t < \xi_1 + \delta, \tag{C.4}$$

and using this inequality we obtain

$$\Delta_{t+1} < (\xi_1 + \delta)\Delta_t. \tag{C.5}$$

The convergence of above recursion gives

$$-\frac{1}{t} \log(\Delta_t) < -\log(\xi_1 + \delta) - \frac{1}{t} \log(\Delta_0). \tag{C.6}$$

Taking the limit on both sides concludes the proof. □

C.1.2 Number of iterations

The total number of iterations, combined with the per-iteration cost, determines the total complexity of an algorithm. Therefore, it is important to derive an upper bound on the total number of iterations of a compositely converging algorithm.

Lemma C.1.2. *For a compositely converging algorithm as in Equation* (C.1) *with coefficients* $\xi_1, \xi_2 \in (0, 1)$*, assume that the initial distance* Δ_0 *satisfies* $\Delta_0 < (1 - \xi_1)/\xi_2$ *and for a given tolerance* ϵ*, define the interval*

$$D = \left(\max \left\{ \epsilon, \frac{\xi_1 \Delta_0}{1 - \xi_2 \Delta_0} \right\}, \Delta_0 \right). \tag{C.7}$$

Then the total number of iterations needed to approximate the true minimizer with ϵ tolerance is upper bounded by $T(\delta_*)$, where

$$\delta_* = \operatorname*{argmin}_{\delta \in D} T(\delta) \tag{C.8}$$

and

$$T(\delta) = \log_2 \left(\frac{\log\left(\xi_1 + \delta\xi_2\right)}{\log\left(\frac{\Delta_0}{\delta}(\xi_1 + \delta\xi_2)\right)} \right) + \frac{\log\left(\frac{\epsilon}{\delta}\right)}{\log(\xi_1 + \xi_2\delta)}. \tag{C.9}$$

Proof of Lemma C.1.2. We have $\Delta_t \to 0$ as $t \to \infty$ by the condition on initial point Δ_0. Let $\delta \in D$ be a real number and t_1 be the last iteration step such that $\Delta_t > \delta$. Then $\forall t \geq t_1$,

$$\begin{aligned} \Delta_{t+1} \leq & \xi_1\Delta_t + \xi_2\Delta_t^2, \\ \leq & \left(\frac{\xi_1}{\delta} + \xi_2\right)\Delta_t^2. \end{aligned} \tag{C.10}$$

Therefore, in this regime, the convergence rate of the algorithm is dominated by a quadratically converging term with coefficient $(\xi_1/\delta + \xi_2)$. The total number of iterations needed to attain a tolerance of δ is upper bounded by

$$t_1 \leq \log_2 \left(\frac{\log\left(\xi_1 + \delta\xi_2\right)}{\log\left(\frac{\Delta_0}{\delta}(\xi_1 + \delta\xi_2)\right)} \right). \tag{C.11}$$

When $\Delta_t < \delta$, namely $t > t_1$, we have

$$\begin{aligned} \Delta_{t+1} \leq & \xi_1\Delta_t + \xi_2\Delta_t^2, \\ \leq & \left(\xi_1 + \xi_2\delta\right)\Delta_t. \end{aligned} \tag{C.12}$$

In this regime, the convergence rate is dominated by a linearly converging term with coefficient $(\xi_1 + \xi_2\delta)$. Therefore, the total number of iterations since t_1 until a tolerance of ϵ is reached can be upper bounded by

$$t_2 \leq \frac{\log\left(\frac{\epsilon}{\delta}\right)}{\log(\xi_1 + \xi_2\delta)}. \tag{C.13}$$

Hence, the total number of iterations needed for a composite algorithm as in Equation (C.1) to reach a tolerance of ϵ is upper bounded by

$$T(\delta) = t_1 + t_2 = \log_2 \left(\frac{\log\left(\xi_1 + \delta\xi_2\right)}{\log\left(\frac{\Delta_0}{\delta}(\xi_1 + \delta\xi_2)\right)} \right) + \frac{\log\left(\frac{\epsilon}{\delta}\right)}{\log(\xi_1 + \xi_2\delta)}. \tag{C.14}$$

The above statement holds for any $\delta \in D$. Therefore, we minimize $T(\delta)$ over the set D. □

C.2 Choosing the step size

In most optimization algorithms, step size plays a crucial role. If the dataset is so large that one cannot try out many values of the step size. In this section, we describe an efficient and adaptive way for this purpose by using the theoretical results derived in the previous sections.

In the proof of Lemma 4.3.1, we observe that the convergence rate of NewSamp is governed by the term

$$\left\|\mathbf{I} - \gamma_t \mathbf{Q}^t H_{[n]}(\tilde{\beta})\right\|_2 \leq \left\|\mathbf{I} - \gamma_t \mathbf{Q}^t H_{[n]}(\hat{\beta}^t)\right\|_2 + \gamma_t \left\|\mathbf{Q}^t\right\|_2 \left\|H_{[n]}(\hat{\beta}^t) - H_{[n]}(\tilde{\beta})\right\|_2 \tag{C.15}$$

where $\mathbf{Q}^t$ is defined as in Algorithm 1. The right hand side of the above equality has a linear dependence on γ_t. We will see later that this term has no effect in choosing the right step size. On the other hand, the first term on the right hand size can be written as,

$$\left\|\mathbf{I} - \gamma_t \mathbf{Q}^t H_{[n]}(\hat{\beta}^t)\right\|_2 = \max\left\{1 - \gamma_t \lambda_{\min}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)), \gamma_t \lambda_{\max}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)) - 1\right\}. \tag{C.16}$$

If we optimize the above quantity over γ_t, we obtain the optimal step size as

$$\gamma_t = \frac{2}{\lambda_{\min}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)) + \lambda_{\max}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t))}. \tag{C.17}$$

It is worth mentioning that for the Newton method where $\mathbf{Q}^t = H_{[n]}(\hat{\beta}^t)^{-1}$, the above quantity is equal to 1.

Since NewSamp does not compute the full Hessian $\mathbf{H}_{[n]}(\hat{\beta}^t)$ (which would take $\mathcal{O}(np^2)$ computation), we will relate the quantity in Equation (C.17) to the first few eigenvalues of $\mathbf{Q}^t$. Therefore, our goal is to relate the eigenvalues of $\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)$ to that of $\mathbf{Q}^t$.

By the Lipschitz continuity of eigenvalues , we write

$$\begin{aligned}\left|1-\lambda_{\max}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t))\right| &\leq \left\|\mathbf{Q}^t\right\|_2 \left\|H_S(\hat{\beta}^t) - H_{[n]}(\hat{\beta}^t)\right\|_2, \\ &= \frac{1}{\lambda_{r+1}^t}\mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right). \end{aligned} \tag{C.18}$$

Similarly, for the minimum eigenvalue, we can write

$$\left|\frac{\lambda_p^t}{\lambda_{r+1}^t} - \lambda_{\min}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t))\right| \leq \frac{1}{\lambda_{r+1}}\mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right). \tag{C.19}$$

One might be temped to use 1 and $\lambda_p^t/\lambda_{r+1}^t$ for the minimum and the maximum eigenvalues of $\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)$, but the optimal values might be slightly different from these values if the sample size is chosen to be small. On the other hand, the eigenvalues λ_{r+1}^t and λ_p^t can be computed with $\mathcal{O}(p^2)$ cost and we already know the order of the error term. That is, one can calculate λ_{r+1}^t and λ_p^t and use the error bounds to correct the estimate.

The eigenvalues of the sample covariance matrix will concentrate around the true values, spreading to be larger for large eigenvalues and smaller for the small eigenvalues. That is, if we will we will overestimate if we estimate λ_1 with λ_1^t. Therefore, if we use 1, we will always underestimate the value of $\lambda_{\max}(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t))$, which, based on Equation (C.18) and Equation C.19, suggests a correction term of $\mathcal{O}\left(\sqrt{\log(p)/|S|}\right)$. Further, the top $r+1$ eigenvalues of $[Q^t]^{-1}$ are close to the eigenvalues of $H_{[n]}(\hat{\beta}^t)$, but shifted upwards if $p/2 > r$. When $p/2 < r$, we see an opposite behavior. Hence, we add or subtract a correction term of order $\mathcal{O}\left(\sqrt{\log(p)/|S|}\right)$ to $\lambda_p^t/\lambda_{r+1}^t$ whether $p/2 > r$ or $p/2 < r$, respectively. The corrected estimators could be written as

$$\begin{aligned}\widehat{\lambda_{\max}}\left(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)\right) &= 1 + \mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right), \\ \widehat{\lambda_{\min}}\left(\mathbf{Q}^t H_{[n]}(\hat{\beta}^t)\right) &= \frac{\lambda_p}{\lambda_{r+1}} + \mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right) \qquad \text{if } p/2 > r, \\ &= \frac{\lambda_p}{\lambda_{r+1}} - \mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right) \qquad \text{if } p/2 < r.\end{aligned}$$

We are more interested in the case where $p/2 > r$. In this case, we suggest the step size

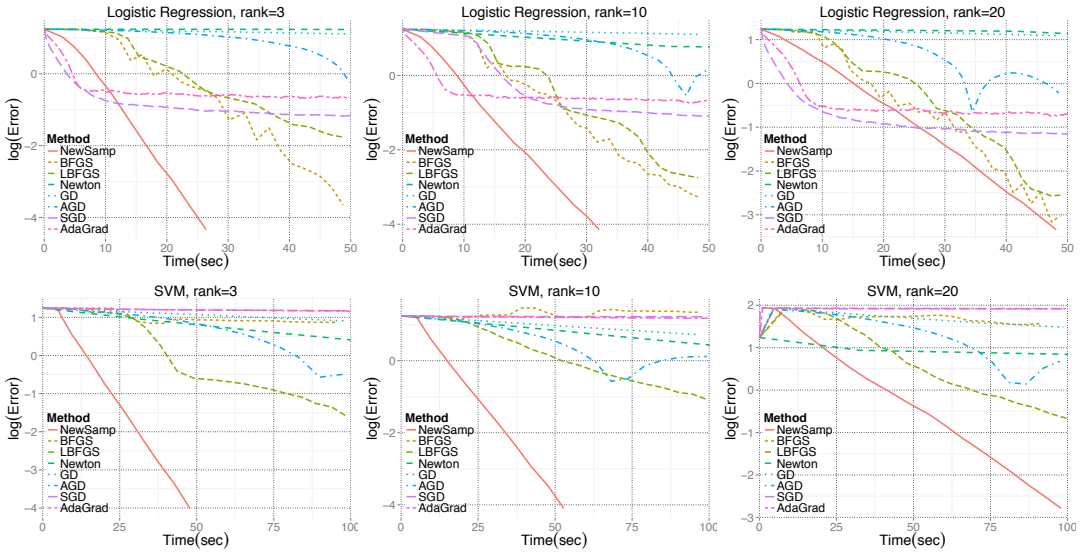

Figure C.1: The plots demonstrate the behavior of several optimization methods on a synthetic data set for training SVMs. The elapsed time in seconds versus log of ℓ_2-distance to the true minimizer is plotted. Red color represents the proposed method NewSamp .

for the iteration step t as

$$\gamma_t = \frac{2}{1 + \frac{\lambda_p^t}{\lambda_{r+1}^t} + \mathcal{O}\left(\sqrt{\frac{\log(p)}{|S|}}\right)} \tag{C.20}$$

which uses the eigenvalues that are already computed to construct $\mathbf{Q}^t$. Contrary to the most algorithms, the optimal step size of NewSamp is generally larger than 1.

C.3 Further experiments and details

In this section, we present the details of the experiments presented in Figure 4.2 and provide additional simulation results.

We first start with additional experiments. The goal of this experiment is to further analyze the effect of rank in the performance of NewSamp . We experimented using r-spiked model for $r = 3, 10, 20$. The case $r = 3$ was already presented in Figure 4.2, which is included in Figure C.1 to ease the comparison. The results are presented in Figures C.1 and the details are summarized in Table C.1. In the case of LR optimization, we observe

Logistic Regression

	Rank=3		Rank=10		Rank=20	
Method	Elapsed(sec)	Iter	Elapsed(sec)	Iter	Elapsed(sec)	Iter
NewSamp	26.412	12	32.059	15	55.995	26
BFGS	50.699	22	54.756	31	56.606	34
LBFGS	103.590	47	64.617	37	107.708	67
Newton	18235.842	449	35533.516	941	31032.893	777
GD	345.025	198	322.671	198	311.946	197
AGD	449.724	233	436.282	272	450.734	290

Support Vector Machines

	Rank=3		Rank=10		Rank=20	
Method	Elapsed(sec)	Iter	Elapsed(sec)	Iter	Elapsed(sec)	Iter
NewSamp	47.755	8	52.767	9	124.989	22
BFGS	13352.254	2439	10672.657	2219	21874.637	4290
LBFGS	326.526	67	218.706	44	275.991	55
Newton	775.191	16	734.480	16	4159.486	106
GD	1512.305	238	1089.413	237	1518.063	269
AGD	1695.44	239	1066.484	238	1874.75	294

Table C.1: Details of the simulations presented in Figures C.1.

through Figure C.1 that stochastic algorithms enjoy fast convergence in the beginning but slows down later as they get close to the true minimizer. The algorithms that come closer to NewSamp in terms of performance are BFGS and LBFGS. Especially when $r = 20$, performance of BFGS and that of NewSamp are similar, yet NewSamp still does better. In the case of SVM optimization, the algorithm that comes closer to NewSamp is Newton method.

We further demonstrate how the algorithm coefficients ξ_1 and ξ_2 between datasets in Figure C.2.

CT Slices Dataset

	LR		SVM	
Method	Elapsed(sec)	Iter	Elapsed(sec)	Iter
NewSamp	9.488	19	22.228	33
BFGS	9.568	38	2094.330	5668
LBFGS	51.919	217	165.261	467
Newton	14.162	5	58.562	25
GD	350.863	2317	1660.190	4828
AGD	176.302	915	1221.392	3635

MSD Dataset

	LR		SVM	
Method	Elapsed(sec)	Iter	Elapsed(sec)	Iter
NewSamp	25.770	38	71.755	49
BFGS	43.537	75	9063.971	6317
LBFGS	81.835	143	429.957	301
Newton	144.121	30	100.375	18
GD	642.523	1129	2875.719	1847
AGD	397.912	701	1327.913	876

Synthetic Dataset

	LR		SVM	
Method	Elapsed(sec)	Iter	Elapsed(sec)	Iter
NewSamp	26.412	12	47.755	8
BFGS	50.699	22	13352.254	2439
LBFGS	103.590	47	326.526	67
Newton	18235.842	449	775.191	16
GD	345.025	198	1512.305	238
AGD	449.724	233	1695.44	239

Table C.2: Details of the experiments presented in Figure 4.2.

C.4 Useful lemmas

Lemma C.4.1. *Let $\mathcal{C}$ be convex and bounded set in $\mathbb{R}^p$ and T_ϵ be an ϵ-net over $\mathcal{C}$. Then,*

$$|T_\epsilon| \leq \left(\frac{\text{diam}(\mathcal{C})}{2\epsilon/\sqrt{p}}\right)^p. \tag{C.21}$$

Proof of Lemma C.4.1. A similar proof appears in [VdVW96]. The set $\mathcal{C}$ can be contained in a p-dimensional cube of size $\text{diam}(\mathcal{C})$. Consider a grid over this cube with mesh width $2\epsilon/\sqrt{p}$. Then $\mathcal{C}$ can be covered with at most $(\text{diam}(\mathcal{C})/(2\epsilon/\sqrt{p}))^p$ many cubes of edge length $2\epsilon/\sqrt{p}$. If ones takes the projection of the centers of such cubes onto $\mathcal{C}$ and considers the circumscribed balls of radius ϵ, we may conclude that $\mathcal{C}$ can be covered with at most

$$\left(\frac{\text{diam}(\mathcal{C})}{2\epsilon/\sqrt{p}}\right)^p \tag{C.22}$$

many balls of radius ϵ. □

Lemma C.4.2 ([GN10])**.** *Let $\mathcal{X}$ be a finite set of Hermitian matrices in $\mathbb{R}^{p\times p}$ where $\forall X_i \in \mathcal{X}$, we have*

$$\mathbb{E}[X_i] = 0, \qquad \|X_i\|_2 \leq \gamma, \qquad \left\|\mathbb{E}[X_i^2]\right\|_2 \leq \sigma^2. \tag{C.23}$$

Given its size, let S denote a uniformly random sample from $\{1, 2, ..., |\mathcal{X}|\}$ with or without replacement. Then we have

$$\mathbb{P}\left(\left\|\frac{1}{|S|}\sum_{i\in S} X_i\right\|_2 > \epsilon\right) \leq 2p\exp\left\{-|S|\min\left(\frac{\epsilon^2}{4\sigma^2}, \frac{\epsilon}{2\gamma}\right)\right\}. \tag{C.24}$$

Lemma C.4.3. *Let Z be a random variable with a density function f and cumulative distribution function F. If $F^C = 1 - F$, then,*

$$\left|\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}]\right| \leq t\mathbb{P}(|Z| > t) + \int_t^\infty \mathbb{P}(|Z| > z)dz. \tag{C.25}$$

Proof. We write,

$$\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}] = \int_t^\infty zf(z)dz + \int_{-\infty}^{-t} zf(z)dz.$$

Using integration by parts, we obtain

$$\begin{aligned}\int zf(z)dz =& - zF^C(z) + \int F^C(z)dz, \\ =& zF(z) - \int F(z)dz.\end{aligned} \tag{C.26}$$

Since $\lim_{z\to\infty} zF^C(z) = \lim_{z\to-\infty} zF(z) = 0$, we have

$$\begin{aligned}\int_t^\infty zf(z)dz =& tF^C(t) + \int_t^\infty F^C(z)dz, \\ \int_{-\infty}^{-t} zf(z)dz =& - tF(-t) - \int_{-\infty}^{-t} F(z)dz, \\ =& - tF(-t) - \int_t^\infty F(-z)dz.\end{aligned} \tag{C.27}$$

Hence, we obtain the following bound,

$$\begin{aligned}\left|\mathbb{E}[Z\mathbb{1}_{\{|Z|>t\}}]\right| =& \left|tF^C(t) + \int_t^\infty F^C(z)dz - tF(-t) - \int_t^\infty F(-z)dz\right|, \\ \leq& t\left(F^C(t) + F(-t)\right) + \left(\int_t^\infty F^C(z) + F(-z)dz\right), \\ \leq& t\mathbb{P}(|Z| > t) + \int_t^\infty \mathbb{P}(|Z| > z)dz.\end{aligned} \tag{C.28}$$

□

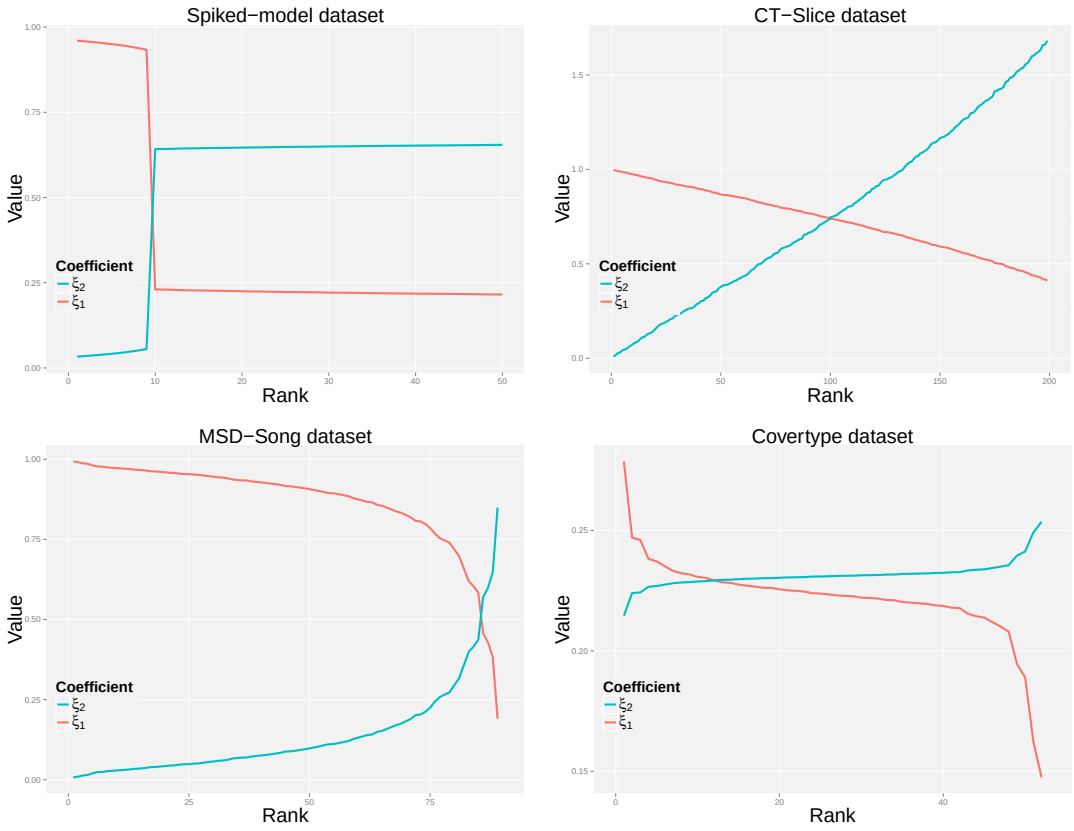

Figure C.2: The plots demonstrate the behavior of ξ_1 and ξ_2 over several datasets.

Bibliography

[Ama98] Shun-Ichi Amari, *Natural gradient works efficiently in learning*, Neural computation **10** (1998), no. 2, 251–276.

[BCNN11] Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal, *On the use of stochastic hessian information in optimization methods for machine learning*, SIAM Journal on Optimization **21** (2011), no. 3, 977–995.

[BD99] Jock A Blackard and Denis J Dean, *Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables*, Computers and electronics in agriculture **24** (1999), no. 3, 131–151.

[BEM13] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari, *Estimating lasso risk and noise level*, Advances in Neural Information Processing Systems, 2013, pp. 944–952.

[BHNS14] Richard H Byrd, SL Hansen, Jorge Nocedal, and Yoram Singer, *A stochastic quasi-newton method for large-scale optimization*, arXiv preprint arXiv:1401.7020 (2014).

[Bis95] Christopher M Bishop, *Neural networks for pattern recognition*, Oxford university press, 1995.

[Bot10] Léon Bottou, *Large-scale machine learning with stochastic gradient descent*, Proceedings of COMPSTAT, Springer, 2010, pp. 177–186.

[Bri82] David R Brillinger, *A generalized linear model with "Gaussian" regressor variables*, A Festschrift For Erich L. Lehmann, CRC Press, 1982, pp. 97–114.

[Bro70] Charles G Broyden, *The convergence of a class of double-rank minimization algorithms 2. the new algorithm*, IMA Journal of Applied Mathematics **6** (1970), no. 3, 222–231.

[BSS05] Andreas Buja, Werner Stuetzle, and Yi Shen, *Loss functions for binary class probability estimation and classification: Structure and applications.*

[BSW14] Pierre Baldi, Peter Sadowski, and Daniel Whiteson, *Searching for exotic particles in high-energy physics with deep learning*, Nature communications **5** (2014).

[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[CCS10] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization **20** (2010), no. 4, 1956–1982.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, *Atomic decomposition by basis pursuit*, SIAM review **43** (2001), no. 1, 129–159.

[CGS10] Louis HY Chen, Larry Goldstein, and Qi-Man Shao, *Normal approximation by stein's method*, Springer Science & Business Media, 2010.

[Cha07] Olivier Chapelle, *Training a support vector machine in the primal*, Neural Computation (2007).

[DE16] Lee H Dicker and Murat A. Erdogdu, *Maximum likelihood for variance estimation in high-dimensional linear models*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, 2016, pp. 159–167.

[DE17] Lee H Dicker and Murat A Erdogdu, *Flexible results for quadratic forms with applications to variance components estimation*, The Annals of Statistics **45** (2017), no. 1, 386–414.

[DGJ13] David L Donoho, Matan Gavish, and Iain M Johnstone, *Optimal shrinkage of eigenvalues in the spiked covariance model*, arXiv preprint arXiv:1311.0851 (2013).

[DHS11] John Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research **12** (2011), no. Jul, 2121–2159.

[DL91] Naihua Duan and Ker-Chau Li, *Slicing regression: a link-free regression method*, The Annals of Statistics (1991), 505–530.

[DLFU13] Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar, *New subsampling algorithms for fast least squares regression*, Advances in Neural Information Processing Systems, 2013, pp. 360–368.

[DM77] John E Dennis, Jr and Jorge J Moré, *Quasi-newton methods, motivation and theory*, SIAM review **19** (1977), 46–89.

[DMMS11] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Sarlòs, *Faster least squares approximation*, Numer. Math. **117** (2011), no. 2.

[EBD16a] Murat A. Erdogdu, Mohsen Bayati, and Lee H. Dicker, *Scalable approximations for generalized linear problems*, arXiv preprint arXiv:1611.06686 (2016).

[EBD16b] Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker, *Scaled least squares estimator for glms in large-scale problems*, Advances in Neural Information Processing Systems, 2016, pp. 3324–3332.

[EF15] Murat A Erdogdu and Nadia Fawaz, *Privacy-utility trade-off under continual observation*, Information Theory (ISIT), 2015 IEEE International Symposium on, IEEE, 2015, pp. 1801–1805.

[EFM15] Murat A Erdogdu, Nadia Fawaz, and Andrea Montanari, *Privacy-utility trade-off for time-series with application to smart-meter data*, Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[EM15] Murat A Erdogdu and Andrea Montanari, *Convergence rates of sub-sampled newton methods*, Advances in Neural Information Processing Systems, 2015, pp. 3052–3060.

[Erd15] Murat A Erdogdu, *Newton-stein method: A second order method for glms via stein's lemma*, Advances in Neural Information Processing Systems, 2015, pp. 1216–1224.

[Erd16] Murat A. Erdogdu, *Newton-stein method: An optimization method for glms via stein's lemma*, Journal of Machine Learning Research **17** (2016), no. 216, 1–52.

[Erd17] Murat A Erdogdu, *Generalized hessian approximations via stein's lemma for constrained minimization*, The Information Theory and Applications (2017).

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of statistical software **33** (2010), no. 1, 1.

[Fis36] Ronald A. Fisher, *The use of multiple measurements in taxonomic problems*, Annals Eugenic **7** (1936), 179–188.

[Fle70] Roger Fletcher, *A new approach to variable metric algorithms*, The computer journal **13** (1970), no. 3, 317–322.

[FS12] Michael P Friedlander and Mark Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM Journal on Scientific Computing **34** (2012), no. 3, A1380–A1405.

[GD14] Matan Gavish and David L Donoho, *Optimal shrinkage of singular values*, arXiv:1405.7511 (2014).

[GKS+11] Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro, *2d image registration in ct images using radial image descriptors*, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011 (2011), 607–614.

[GN10] David Gross and Vincent Nesme, *Note on sampling without replacing from a finite collection of matrices*, arXiv preprint arXiv:1001.2738 (2010).

[GNS09] Igor Griva, Stephen G Nash, and Ariela Sofer, *Linear and nonlinear optimization*, SIAM, 2009.

[Gol70] Donald Goldfarb, *A family of variable-metric methods derived by variational means*, Mathematics of computation **24** (1970), no. 109, 23–26.

[Gol07] Larry Goldstein, l^1 *bounds in normal approximation*, Annals Probability **35** (2007), 1888–1930.

[GR97] Larry Goldstein and Gesine Reinert, *Stein's method and the zero bias transformation with application to simple random sampling*, Annals of Applied Probability **7** (1997), 935–952.

[Gro11] David Gross, *Recovering low-rank matrices from few coefficients in any basis*, Information Theory, IEEE Transactions on **57** (2011), no. 3, 1548–1566.

[GS02] Alison L Gibbs and Francis Edward Su, *On choosing and bounding probability metrics*, ISR **70** (2002), no. 3, 419–435.

[HJS01] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny, *Direct estimation of the index coefficient in a single-index model*, Annals of Statistics (2001), 595–623.

[HK70] Arthur E Hoerl and Robert W Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[HKW99] David P Helmbold, Jyrki Kivinen, and Manfred K Warmuth, *Relative loss bounds for single neurons*, IEEE Transactions on Neural Networks **10** (1999), no. 6, 1291–1304.

[HL93] Peter Hall and Ker-Chau Li, *On almost linearity of low dimensional projections from high dimensional data*, The annals of Statistics (1993), 867–889.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, no. 2, 217–288.

[HS52] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand. **49** (1952), 409–436.

[KD05] S Sathiya Keerthi and Dennis DeCoste, *A modified finite newton method for fast solution of large scale linear svms*, Journal of Machine Learning Research, 2005, pp. 341–361.

[KEO15] Ritesh Kolte, Murat A Erdogdu, and Ayfer Ozgur, *Accelerating svrg via second-order information*, NIPS Workshop on Optimization for Machine Learning, 2015.

[KF09] Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT press, 2009.

[KKSK11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai, *Efficient learning of generalized linear and single index models with isotonic regression*, Advances in Neural Information Processing Systems, 2011, pp. 927–935.

[KS09] Adam Tauman Kalai and Ravi Sastry, *The isotron algorithm: High-dimensional isotonic regression.*, Proceedings of the 22nd Annual Conference on Learning Theory, 2009.

[LD89] Ker-Chau Li and Naihua Duan, *Regression analysis under link violation*, The Annals of Statistics (1989), 1009–1052.

[LD09] Bing Li and Yuexiao Dong, *Dimension reduction for nonelliptically distributed predictors*, The Annals of Statistics (2009), 1272–1298.

[Li91] Ker-Chau Li, *Sliced inverse regression for dimension reduction*, Journal of the American Statistical Association **86** (1991), no. 414, 316–327.

[Lic13] Moshe Lichman, *UCI machine learning repository*, 2013.

[LRF10] Nicolas Le Roux and Andrew W Fitzgibbon, *A fast natural newton method*, Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 623–630.

[LRMB08] Nicolas Le Roux, Pierre-Antoine Manzagol, and Yoshua Bengio, *Topmoumoute online natural gradient algorithm*, Advances in neural information processing systems, 2008, pp. 849–856.

[LSS14] Jason D Lee, Yuekai Sun, and Michael A Saunders, *Proximal newton-type methods for minimizing composite functions*, SIAM Journal on Optimization **24** (2014), no. 3, 1420–1443.

[LWK08] Chih-J Lin, Ruby C Weng, and Sathiya Keerthi, *Trust region newton method for logistic regression*, JMLR (2008).

[Mar10] James Martens, *Deep learning via hessian-free optimization*, Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 735–742.

[MEWL11] Thierry B. Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere, *The million song dataset*, ISMIR-11, 2011.

[MJC+14] Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, Joel A Tropp, et al., *Matrix concentration inequalities via the method of exchangeable pairs*, The Annals of Probability **42** (2014), no. 3, 906–945.

[MN89] Peter McCullagh and John A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, 1989.

[NB72] John A Nelder and R. Jacob Baker, *Generalized linear models*, Wiley Online Library, 1972.

[Nes83] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Dokl. **27** (1983), 372–376.

[Nes13] Yurii Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.

[PS75] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM Journal of Numerical Analysis **12** (1975), 617–629.

[PV16] Yaniv Plan and Roman Vershynin, *The generalized lasso with non-linear observations*, IEEE Transactions on information theory **62** (2016), no. 3, 1528–1537.

[PW15] Mert Pilanci and Martin J Wainwright, *Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence*, arXiv preprint arXiv:1505.02250 (2015).

[RKM16] Farbod Roosta-Khorasani and Michael W Mahoney, *Sub-sampled newton methods i: globally convergent algorithms*, arXiv preprint arXiv:1601.04737 (2016).

[RM51] Herbert Robbins and Sutton Monro, *A stochastic approximation method*, Annals of mathematical statistics (1951).

[RT08] Vladimir Rokhlin and Mark Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proceedings of the National Academy of Sciences **105** (2008), no. 36, 13212–13217.

[RW10] Mark D Reid and Robert C Williamson, *Composite binary losses*, Journal of Machine Learning Research **11** (2010), no. Sep, 2387–2422.

[Sch89] Mark J Schervish, *A general method for comparing probability assessors*, The Annals of Statistics (1989), 1856–1879.

[SDPG13] Jascha Sohl-Dickstein, Ben Poole, and Surya Ganguli, *An adaptive low dimensional quasi-newton sum of functions optimizer*, arXiv preprint arXiv:1311.2115 (2013).

[Sha70] David F Shanno, *Conditioning of quasi-newton methods for function minimization*, Mathematics of computation **24** (1970), no. 111, 647–656.

[SHY+13] Andrew Senior, Georg Heigold, Ke Yang, et al., *An empirical study of learning rates in deep neural networks for speech recognition*, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 6724–6728.

[SLRB] Mark Schmidt, Nicolas Le Roux, and Francis Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, 1–30.

[SS02] Bernhard Schölkopf and Alexander J Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[Ste81] Charles M Stein, *Estimation of the mean of a multivariate normal distribution*, Annals of Statistics (1981), 1135–1151.

[Sti81] Stephen M Stigler, *Gauss and the invention of least squares*, The Annals of Statistics (1981), 465–474.

[SYG07] Nicol Schraudolph, Jin Yu, and Simon Günter, *A stochastic quasi-newton method for online convex optimization.*

[TAH15] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Lasso with non-linear measurements is equivalent to one with linear measurements*, Advances in Neural Information Processing Systems, 2015, pp. 3420–3428.

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[Tro12] Joel A Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics (2012).

[Vap98] Vladimir Vapnik, *Statistical learning theory*, vol. 2, Wiley New York, 1998.

[VdV00] Aad W Van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 2000.

[VdVW96] Aad W Van der Vaart and Jon A Wellner, *Weak convergence*, Springer, 1996.

[Ver10] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, 2010, arXiv:1011.3027.

[Ver12] ———, *How close is the sample covariance matrix to the actual covariance matrix?*, Journal of Theoretical Probability **25** (2012), no. 3, 655–686.

[VP12] Oriol Vinyals and Daniel Povey, *Krylov subspace descent for deep learning*, International Conference on Artificial Intelligence and Statistics, 2012, pp. 1261–1268.

[WG11] Kristian Woodsend and Jacek Gondzio, *Exploiting separability in large-scale linear support vector machine training*, Computational Optimization and Applications **49** (2011), no. 2, 241–269.

[WJ08] Martin J Wainwright and Michael I Jordan, *Graphical models, exponential families, and variational inference*, Foundations and Trends in Machine Learning **1** (2008), 1–305.

[ZH05] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.